Determination of the genetic cause of an internationally unique, naturally occurring muscular dystrophy in Western Australian Merino sheep

This thesis is presented for the Honours degree in Biomedical Science at Murdoch University

Jez Supreme BSc. Molecular Biology BSc. Biotechnology

11/1/2013

Supervisors:

A/Prof Kim Carter
Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, Perth, Australia

A/Prof Kristen Nowak
Western Australian Institute for Medical Research, Centre for Medical Research, University of Western Australia, Perth, Australia

A/Prof Wayne Greene
Murdoch University, Perth, Australia

Prof. Nigel Laing
Western Australian Institute for Medical Research, Centre for Medical Research, University of Western Australia, Perth, Australia

Declaration Page

I declare this thesis is my own account of my research and contains as its main content, work which has not previously been submitted for a degree at any tertiary educational institution.

.....

Jez Supreme

Abstract

Muscular dystrophies (MDs) are neuromuscular disorders characterised by chronic, usually progressive, skeletal muscle weakness. Individuals often lose walking ability and can suffer terminal cardiorespiratory complications. Determining the genetics of a disease helps provide diagnosis, prognosis, genetic counselling, and the basis for rational therapeutic design.

A naturally occurring sheep model of autosomal recessive congenital MD was identified in WA in the 1950's and preserved as a research colony. The pathological features and distribution of this MD is novel. A sheep model of MD is incredibly valuable; sheep have similar skeletal muscle mass to humans, representing a significant improvement over smaller mammals in which to trial therapies. Successfully characterising the causative gene(s) would enable a possible target for new therapies and may open new lines of investigation into better understanding and treatment of MD in humans.

This project utilised a two-pronged approach to investigate the genetics of this ovine MD. First, bioinformatics analysis of SNP genotyping for multiple individuals in the flock by a 50,000 SNP array in combination with the latest sheep genome reference build released by the International Sheep Genome Consortium, enabling homozygosity mapping, genetic linkage and association mapping. Second, molecular biological approaches further explored the identified prime candidate gene by cDNA sequencing.

This research project identified *ROCK2* as the prime candidate gene most likely harbouring a mutation causing the muscular dystrophy in this internationally unique ovine model. It also demonstrated for the first time in sheep the existence of *ROCK2m,* an isoform of *ROCK2* preferentially expressed in skeletal muscle. This work has set the stage for further investigations into *ROCK2m* and the ovine MD which will hopefully pinpoint the causative disease mutation.

Table of Contents

Acknowledgements

This thesis would not have been possible without the support, care and guidance of myriad people. First, I must gratefully acknowledge the expert knowledge, insight, encouragement, and hours of patient teaching provided by my supervisors A/Prof Kim Carter, A/Prof Kristen Nowak, and Prof. Nigel Laing (and thanks for handling the Murdoch paperwork, A/Prof. Wayne Greene!).

With no experience in bioinformatics beyond being a videogame geek and thinking it was a cool area, Kim took me under his wing and spent FAR more time than he had to spare guiding me through the fundamentals of UNIX, bioinformatics methodologies and scripting in Perl. It might be awhile before he takes on another biology major for a bioinformatics project, but I am extremely grateful he wasted his time with me. Non-model organisms, eh?!

To Kristen, who seemed to be available and enthusiastic so much more than a human could be capable of: THANK YOU. Without your consistent enthusiasm and expert advice on everything moleculr biology-related this project could never have begun (and I certainly couldn't have completed it). Without you and Kim I'm not sure my scientific career would be happening either. You guys rock.

Nigel, you are a neuromuscular disease Jedi; thanks for stooping to deal with such an ignorant Padawan.

To everyone in S203: Dr. Richard Francis (you're a doctor now, right?), Denise Anderson, and Matt Cooper, thank you so much for being so welcoming and helping me when I needed it. Thanks especially to Richard for being my backup bioinformatician whenever Kim was unavailable. You guys made TICHR way more fun than it otherwise would have been.

To Royston Ong & Elyshia McNamara: thank you both SO MUCH for your help. You guys both spent a LOT of time and effort helping me get the Mol. Bio. side of things running and disabusing my ignorance. Thanks also to Kyle Yau for proving once and for all that the linkage problems weren't my fault. Thanks too to Dr. Rachael Duff for your sequence analysis expertise, and to everyone else at WAIMR who helped me out.

Finally, thank you Nicole. Thank you for Egg.

Table of Figures

List of Tables

Definitions & Abbreviations

Allele Alternative DNA sequence at a locus

Amino acid Individual components of a protein. Dictated by a 3 base codon of DNA.

DNA Deoxyribonucleic acid

Exon Segment of a gene which is translated to an end product. A given gene may have multiple exons

Exon splicing Process by which the exons of a gene are joined together as a step towards the production of a mature RNA product. May have alternative forms in which exons are absent, introns preserved, or the inclusion of an alternative acceptor site changes the upstream or downstream exon

Gene A section of DNA that gets transcribed into RNA and then into protein

Genetic locus A particular point in the genome

Genetic Polymorphism A DNA variant occurring at an incidence of less than 99%

GWAS Genome-Wide Association Study

Intron Segment of a gene which is not translated into a product. A given gene may have multiple introns.

Map Distance/Unit usually expressed as either map unit (MU) or centimorgan (cM) is a measure of recombination frequency between two genetic loci. One cM represents the genetic distance between genes at which one product of meiosis in 100 will lead to a recombinant product (recombination frequency of 1%).

Moore's Law Observation that over the history of computing hardware the speed of processors approximately doubles every 2 years; with a concomitant drop in cost

MyD Myotonic Dystrophy

MD Muscular Dystrophy

NM Nemaline myopathy

NMD NeuroMuscular Disorders

Nucleotide A single base of DNA. DNA is made up of four nucleotide bases; A, T, G, C

1. Introduction and Literature Review

1.1. Overview

This literature review gives an outline of genetics and disease, genomics and associated bioinformatics challenges, neuromuscular disorders, the ovine congenital progressive muscular dystrophy model, comparisons of the model with human disease and the field of bioinformatics. The section concludes with the research hypothesis and aims, and the significance of this project work. The following is a short overview of this honours project and its relevance.

Muscular dystrophies (MDs) are neuromuscular disorders, genetic in nature, which present with chronic, usually progressive, weakness of skeletal muscle (used for voluntary movements). In affected individuals, MD negatively affects the performance of daily activities; in many cases leading to the loss of walking ability and complications resulting in early death. Congenital MDs (CMDs) are usually present from birth and represent life-long disease, with a high burden placed on affected individuals and caregivers. The selective targeting of muscle in MD is not yet well understood, but animal models can be utilised to explore disease pathology and progression, and to trial potential treatments of palliative or therapeutic nature.

This project utilises bioinformatics tools and techniques to identify the cause of a naturally occurring MD in sheep, in the hope of this animal model being useful in the research of human disease. The similarities between sheep skeletal muscle mass and that of humans, in contrast to the differences between that of humans and the most popular current animal models, such as mice, will allow a much more effective translation of trial treatment modalities to medical intervention. The greater skeletal muscle mass of sheep also provides improved opportunity to examine affected tissue, important in developing pathophysiological understanding.

Through use of genetic homozygosity mapping and association analysis, assisted by linkage analysis (all discussed later in this section) this project aims to identify genomic regions associated with development, or risk of development, of this disease. Through such analyses we hope to gain a greater understanding of not only this particular disease model, but of how MD selectively affects skeletal muscle tissue and of possible mechanisms of treatment. After identification of a prime gene candidate which was significant both statistically and biologically, this candidate was further explored using

molecular biological techniques to provide additional evidence for or against its involvement in the pathology.

1.2. Genetics & Disease

1.2.1. Importance of Genetics – could be removed or drastically reduced

Our genome provides the template for our growth and cellular differentiation, dictating to a large extent the resistances and susceptibilities of our lifetime (Pierce, 2010). This complete collection of our DNA; the instructions for building and maintaining ourselves as an organism, is organised into 23 pairs of homologous chromosomes in humans (Francomano and Kazazian, 1986). The inherited genome, through interaction with the environment, plays a major role in determining our ultimate fate (Hunter, 2005). Inheritance of a random recombination of our parents' respective genomes through sexual reproduction results in the vast majority of our genes working correctly, but also of genetic variants with the potential to be disease-causing, or which are otherwise abnormal (Pierce, 2010, Tenesa and Haley, 2013). A functional unit of DNA is referred to as a gene and can be conceptualised as a single program written to create a specific protein product (Pierce, 2010).

1.2.2. DNA to Protein

Protein is the major structural and functional component of organisms (Pierce, 2010). Functional protein in eukaryotes is generated from DNA by a multipart process, the overall sequence of events indicated graphically (Fig. 1.1.); first being transcribed into pre-mRNA and then processed, with the addition of a 5' cap, poly-A tail, and exons spliced into mature mRNA (Clark, 2005).

Alternate splicing (Fig. 1.2), can generate an alternative isoform of the product, which may have distinctly different biological activity than that of the more common isoform (Clark, 2005). The base sequences at the splice junctions are of major significance to the correct functioning of this mechanism and a significant fraction of

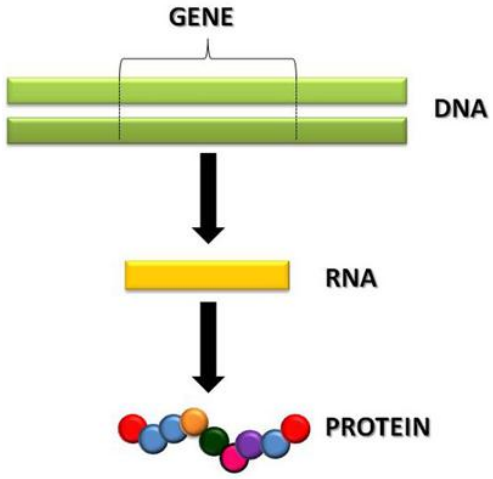


Figure 1.1. The central dogma of molecular biology. DNA is transcribed into RNA and then translated into protein. Figure from (Ianello Giassetti et al., 2013)

human disease is the result of genetic mutations at these sites (López-Bigas et al., 2005, Crotti and Horowitz, 2009). Alternative splicing has been recently reported to take place in up to 94% of eukaryotic multi-exon genes, demonstrating that gene expression is not a simple instruction-to-product process and can be variable even in the case of a single gene, underscoring the importance of alternative splicing in understanding gene expression (reviewed by Chen et al., 2012).

Subsequent to transcription, mature mRNA undergoes translation in the ribosome, first forming the ribosome-mRNA complex and being read by the ribosome as triplet codons, each triplet encoding an amino acid through the redundant genetic code. The correct amino acid is added to the sequence as transfer RNA is brought to complex and matches anticodon to codon for each triplet (Clark, 2005). The completed sequence of amino acids is referred to as a protein and, subsequent to the process of protein processing, this becomes a functional unit within the organism (Becker et al., 2008). This protein may be an enzyme, structural component or hormone and has a biological effect on the organism (Clark, 2005).

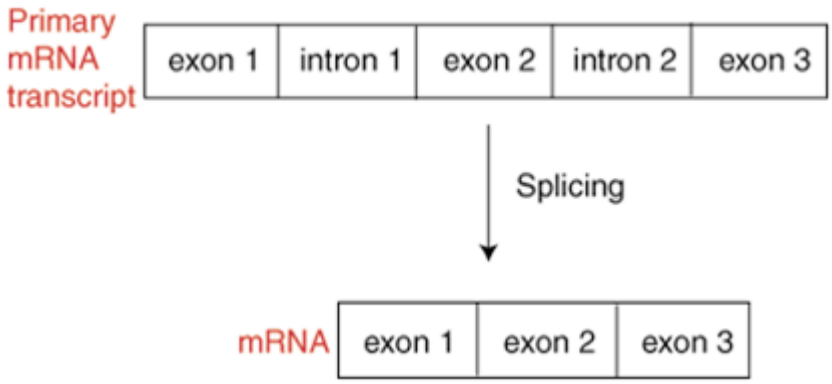


Figure 1.2. Exon splicing. Introns are removed from the pre-mRNA and exons joined. Alternative splicing can generate different isoforms through multiple mechanisms. (Figure from Sparknotes, 2013)

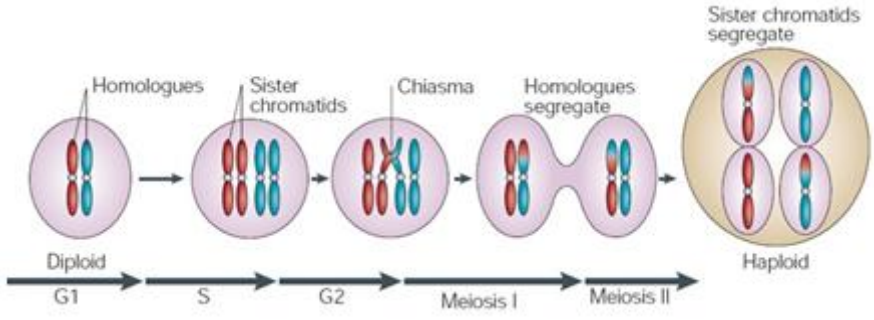


Figure 1.3. Meiosis in diploid organisms. (Figure adapted from Marston and Amon, 2004)

The human genome is diploid; each chromosome paired to its homologous partner and inherited one from each parent (Becker et al., 2008). For each gene there are two copies, or alleles, the DNA sequence of which may differ between copies (Clark, 2005). The generation of haploid DNA through meiosis (Fig. 1.3), in which only one copy of each chromosomal pair is present, is the basis by which sexual recombination of parental DNA takes place (Pierce, 2010). This haploid cell is referred to as a gamete, and the combination of both a maternal and paternal gamete results in the genome of a new diploid offspring (Pierce, 2010).

The generation of haploid DNA for the purpose of reproduction is prone to error through both the copying of DNA by DNA polymerase, which has an error rate of approximately 1 in every 100 000 nucleotides, but also of errors related to recombination processes (Tempest, 2011). While there are mechanisms in place to recognize and correct these, they are not infallible and spontaneous genetic changes can result in deleterious traits in offspring (Tempest, 2011). Where these negative mutations are passed on to offspring we observe the generation of a new heritable disease trait.

1.2.3. Genetic Disease

To some extent, all disease is genetic in nature, whether the breakdown of basic biological functions or the interplay of a host genome with a pathogen. Disease represents non-functional or aberrant behaviour of specific functions or tissues within an organism (Becker et al., 2008). Though all disease can be said to at least to have a genetic component, not all genetic mutations lead to disease (Pierce, 2010). Where protein production or gene function is unaffected by a mutation in a non-coding, non-regulatory region, or a synonymous region undergoes mutation but the resulting amino acid does not change, the mutation is referred to as 'silent', though there is evidence that even translationally silent mutations can result in biological change (Cartegni et al., 2002, Mullard, 2007, Pierce, 2010).

At other times, a mutation may occur in such a genomic region as to have detrimental effects on the organism; this may include disruption of protein formation, aberrant regulation of genetic expression, changes to exon splicing, or a host of other possibilities (Pierce, 2010). These effects can range from the relatively benign, to life-long disease, to spontaneous termination of an affected fetus (Zatkova et al., 2004, Pierce, 2010).

Mutations may be a change to a single base, or they may affect large sequences of DNA (Pierce, 2010). A point mutation affects only one base, while a frameshift mutation can change the reading frame of gene (Pierce, 2010). An insertion or deletion, for instance, changes the resulting amino acid for all subsequent triplet codons (Pierce, 2010). DNA copy number variants (CNVs) in which a simple sequence spanning at least 1000 bases is of variable length between individuals is a contributor to disease, having been associated with a number in recent years; such as Alzheimer's disease, Crohn's disease and myotonic dystrophy (Rovelet-Lecrux et al., 2006, Fellermann et al., 2006, Ashizawa and Sarkar, 2011). Gross changes to the genome through mutation may result in large-scale deletions of chromosomal segments, leading to loss of function from affected genes and downstream effects (Pierce, 2010).

Genetic mutations can be caused by a wide range of factors; copy error, chemical mutagens, retroviral infections, or gene duplications and deletions resulting from the errors in recombination of haploid DNA during sexual reproduction are all possible mechanisms (Becker et al., 2008, Pierce, 2010). These accumulated errors become heritable and for any given lineage these potentially deleterious changes may lead to

disease. Congenital diseases are those that are present from birth, and generally genetic in nature (meaning that they are heritable) but they do not necessarily arise from an inherited disorder (Laing, 2012).

1.2.4. Heritable Disease

Heritable disease can be the result of mutations or dysregulation in which are monogenic, such as cystic fibrosis, which are referred to as Mendelian disorders. It can also be the result of complex polygenic disorders involving many genetic loci and environmental interactions, such as in multiple sclerosis (Kerem et al., 1990, The International Multiple Sclerosis Genetics Consortium, 2010, Tenesa and Haley, 2013). A simple disease gene variant in the diploid human genome may have dominant expression, in which a single inherited allele of the defective gene is enough to manifest the disease phenotype. Expression of recessive disease requires the inheritance of two defective alleles in order for the disease to manifest (Tenesa and Haley, 2013). In many cases, the interaction of the genome with environmental factors also plays a role in determining disease manifestation and/or severity (Tenesa and Haley, 2013).

1.2.5. Disease & Treatment Investigation

Traditionally, the development of understanding related to disease processes, and of treatment modalities, is a multipart process involving diverse areas of the biological sciences and expert collaboration from a range of sub-disciplines.

The first step in understanding disease is the application of basic biological science principles; how do the presented symptoms suggest disease? Following this, the next step is often to undertake an *in vitro* investigation, in which the disease process or potential drug treatment is examined in cell culture (Ong et al., 2013). Recent years have seen the *in silico* modelling of disease and treatment to be an effective method for supplementing these findings prior to taking the research to its next step; subsequently, the disease is examined in a living biological system (Ong et al., 2013). This *in vivo* investigation is generally the most informative means of exploring disease pathogenesis and the safety of potential treatments prior to clinical trials, and it is for this reason that animal models represent a highly valuable tool in disease investigation (Allamand and Campbell, 2000, Perel et al., 2007).

1.2.6. Animal Models of Human Disease

Animal models are used to study disease causes, patterns and treatments by examination of animals with a disease state mimicking that of human disease; this may be naturally occurring or the result of biological manipulation (Washington et al., 2009, Ng et al., 2012, Prather et al., 2013). For these models to be informative it is important to characterise the genotype and involved biological pathways in order to allow ease of investigation, the communication of findings, the generation of new affected individuals for study, and to provide the ability to extrapolate any findings into the understanding of human disease processes, management and treatment (Thyagarajan et al., 2003, Washington et al., 2009, van der Worp et al., 2010, Prather et al., 2013). The conclusions drawn from animal models will invariably contain inconsistencies as compared to humans but this form of biological interrogation allows a far more complete investigation than *in vitro* approaches and in a much more ethically permissive environment than in humans (van der Worp et al., 2010). The use of *in silico* models to study disease patterns and possible drug treatments provides a useful synergy, allowing development of further understanding without the death of animals (Ranganatha, 2012).

1.3. Genomics & Bioinformatics Challenges

1.3.1. Genomics

Genomics as a field has been developing at an exponential pace over the last two decades, brought about by rapid increases in technological power and computing complexity, in combination with advances in gene sequencing methodology and technologies (Magi et al., 2010, Alkan et al., 2011). It is the investigation, annotation and characterisation of the genetic code, providing us with the ability to intelligently interrogate the code of life in order to better understand its processes (Hawkins et al., 2010).

The beginnings of genomics owes not a small debt to the passage of legislation allowing research into the effects of ionising radiation on biological molecules and the creation of radioisotopes (U.S. Congress, 1974). These developments opened the door for Sanger sequencing of molecules and, in combination with the goals of the human genome project for reducing the costs of sequencing and development of automated

sequencing technologies, enabled large-scale investigations of life's basic code (U.S. Department of Energy, 1990).

The story of the development of high-throughput genomics is really the story of the Human Genome Project. This has been the fundamental driver of associated technologies and a lightning rod for vast amounts of funding poured into the field (Collins et al., 2003). While the human genome initiative was announced in 1986, the 15-year project formally began in 1990. Over the next 11 years rapidly advancing technological and theoretical developments such as the use of polymerase chain reaction (PCR), bacterial artificial chromosomes (BACs), shotgun sequencing and the development of oligonucleotide microarrays led to the confirmation of the existence of a third form of life (the Archaea), the publishing of complete DNA sequences of multiple smaller genomes, the publication of the initial working draft of the full human genome sequence in 2000 and 99% of the euchromatic genome completed by 2004 (the 50th anniversary of Watson & Crick's discovery of the structure of DNA itself) (Woese and Fox, 1977, Venter et al., 2001, Human Genome Consortium, , 2004).

The development of PCR allowed the amplification of DNA samples to the point they could be analysed more readily; by amplifying target sequences we could then examine specific genomic regions of interest (Bartlett and Stirling, 2003). These provided the first real mechanism of genomic investigation for evidence of genetic linkage to disease, especially in notable family studies localising complex disorders such as autoimmune disease (Altmüller et al., 2001).

The development of bacterial artificial chromosomes represented a key advance in genomics; they are propagated in *Escherichia coli*, which are able to carry large (~150 kbp) inserts stably (Venter et al., 1998). These BACs can then each be individually sequenced. This technique can be used to target specific regions of DNA or randomly, and the resulting sequences placed into a developed genomic map.

The process of breaking up the DNA molecule into smaller fragments and sequencing them concurrently is referred to as 'shotgun sequencing' and the basic approach was first developed shortly after that of Sanger sequencing itself (Sanger et al., 1982). The present-day approach uses computational analysis to reassemble the resulting fragments into coherent sequences referred to as 'contigs' (contiguous segments) and utilises paired end reads to build 'scaffolds' over larger sections in order

to mitigate the difficulties in placing these correctly, especially where DNA sequences are highly repetitive such as in CNVs (Waterston et al., 2002). This approach may be used in two major ways; the first is hierarchical shotgun assembly, in which the genome is broken up into an overlapping collection of intermediate clones like bacterial artificial chromosomes, the sequence of the individual intermediates then analysed and the BAC sequences merged back together (Waterston et al., 2002). The second approach is whole-genome shotgun assembly, in which the entire genome is fragmented and sequenced concurrently, though there is greater risk in reorganising these fragments without the framework or a BAC library, paired-end sequencing largely overcomes this drawback (reviewed by Shendure and Ji, 2008). At the time the human genome project was concluding, it was suggested that the most effective approach would be to use whole-genome shotgun sequencing and fill in remaining gaps through targeted sequencing using BACs (Venter et al., 1998).

The development of oligonucleotide microarrays allowed the identification of genomic variation, especially single nucleotide polymorphisms and copy number variation, far more cost effectively and easily than DNA sequencing (Maskos and Southern, 1992, Sebat et al., 2004). As genetic markers, copy number variations are easily identified, highly polymorphic and can be mapped throughout the genome (Sebat et al., 2004, Levy et al., 2007). Microarrays have been adopted throughout the biological sciences, being used for disease gene discovery, expression profiling and across multiple species (Meltzer, 2001, Conway and K, 2003, Vallee et al., 2006).

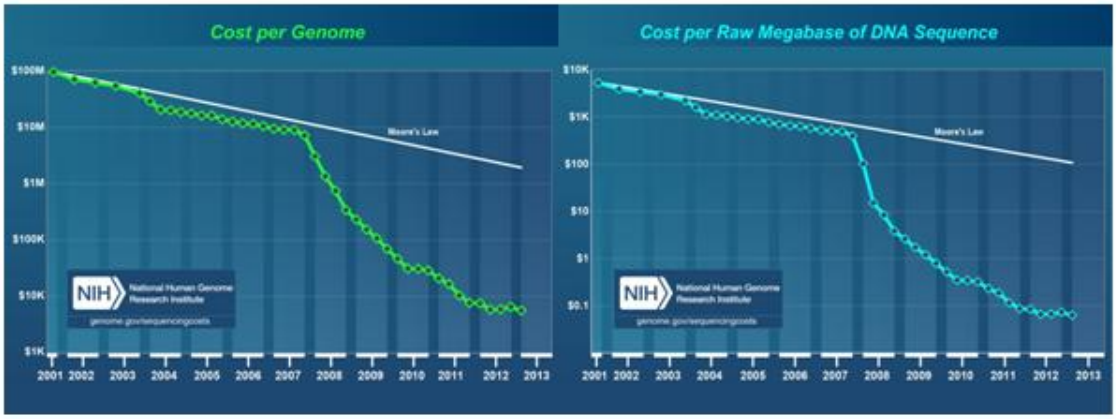


Figure 1.4. Change in DNA sequencing costs over time. (Figures from Wetterstrand, 2013)

The years since the project's formal conclusion have only seen a more rapid increase in the speed and economy of DNA sequencing and genomics, accelerating even faster than Moore's law would have predicted (Fig. 1.4). While formally concluded, the

project is ongoing; the scientific community continues to explore the specific functions of genes and their variants, the true complexity of the human genome remains to be understood (Frazer, 2012).

The genomes of many organisms have been well characterised over recent years; the human mitochondrial genome (16.6 kb) sequenced in 1981, the first eukaryotic chromosome (chr 3 of *Saccharomyces cerevisiae*, 315 kb) sequenced in 1992, the first free-living organism (*Haemophilus influenza*, 1.8 Mb) sequenced in 1995, the first complete eukaryotic genome sequenced in 1996 (*S. cerevisiae*, 12.1 Mb), and the human genome being completed in 2006 (3.2 Gb) (Anderson et al., 1981, Oliver et al., 1992, Fleischmann et al., 1995, Goffeau et al., 1996, Dhand, 2006). As of June, 2013 the complete genome of 3299 viruses, 2497 archaea and bacteria, and 190 eukaryotes are available to the research community (Kanehisa and Goto, 2000, Kanehisa et al., 2012, EMBL-EBI, 2013).

1.3.2. Traditional and Second-Generation DNA Sequencing

The science of sequencing genomes has changed a great deal over the years; from the development of Sanger sequencing in 1977 the field has grown by leaps and bounds along with computing power and more advanced technologies (Sanger et al., 1977, Hawkins et al., 2010). Next Generation Sequencing (NGS) (otherwise referred to as Second-Generation Sequencing) such as Illumina GAII and Roche454 are much more cost effective and speedy than the first generation technologies, and now Third Generation Sequencing (TGS) technologies are in development, with the aim of further reducing costs and increasing the accuracy of genotyping (Check Hayden, 2009, Schadt et al., 2010).

1.3.3. Bioinformatics Challenges for NGS

Sequencing using next-generation technologies has dropped dramatically in price, but the analysis of the results remains a difficult challenge; while advancements in the field are leading to the $1000 sequenced genome, the analysis represents a process that is lengthy by comparison and with costs that are considerably greater, and often hard to estimate (Service, 2006, Sboner et al., 2011). Many aspects of the data processing, tools investigation and use, problem solving and optimization, in addition to the human hours (Sboner et al., 2011) required are difficult to quantify, giving rise to the adage ‘it’s the era of the $1000 genome, but the $100,000 analysis’ (Mardis, 2010). The inability to automate much of the work represents a major bottleneck and, while sequencing costs

may have dropped more quickly than Moore's law, storage capacity has only kept pace, meaning that this, too, represents a bottleneck for effective use of NGS for genomic discovery (Sboner et al., 2011).

NGS technologies have a wide range of applications; used for variant discovery by resequencing target regions of interest or whole genomes, de novo assemblies, transcriptome analysis using RNA-sequencing (RNA-seq), genome-wide profiling of epigenetic markers, and species classification with gene discovery in metagenomics studies (reviewed by Metzker, 2010). To gain an informative result from a mass of NGS reads is computationally intensive and requires a multistep pipeline (Nielsen et al., 2011, Dolled-Filhart et al., 2013).

The process of mapping millions of short reads to the billions of possible positions within the genome is not a computationally trivial step; specialised software must consider the likely start point for each read and account for variation in base call quality (Horner et al., 2010). This is achieved by use of alignment algorithms that have been designed to deal with specialised input, meaning that the parameters of this are subject to stringent criteria. The length of reads, platform of generation, whether the reads are single- or paired-end, whether the resulting alignment can have gaps and the number of individual reads all factor in to what is the most appropriately designed algorithm for alignment of the reads to the reference genome (Horner et al., 2010). Many algorithms focus on speed of alignment over quality, while some take significantly longer in the pursuit of quality, and still others try to blend the two approaches. There is no 'gold standard' alignment algorithm, and the field is rapidly shifting as unserviced niches are discovered and tools written to fill them (Horner et al., 2010).

In most cases, the tools developed to meet these niches are written for use in the Linux/UNIX environments, meaning that implementation involves adept manipulation of the command line and preferably a familiarity with scripting and programming (Stajich and Lapp, 2006). The quickly changing nature of tools and platforms means that there is rarely time for individual tools to mature to the point that they are ported to additional platforms or have all outstanding issues with compatibility and use resolved (Gonzalez-Galarza et al., 2012). Much of the work is developed through the open-source paradigm, meaning that these tools are free and the source code available for modification, but there are inconsistent levels of support and documentation, and the

output of a specific software tool may at times require considerable expertise to parse (Stajich and Lapp, 2006, Kumar and Dudley, 2007).

In DNA sequencing (DNA-seq) using NGS technologies, it is subsequent to the alignment of reads to the corresponding reference genome that variant calling is undertaken (Horner et al., 2010). The major computational challenge at this step is to distinguish 'true' variants from errors in alignment or in base-calling by the sequencer platform, and generally requires a database of known polymorphisms in order to gain good results (Nielsen et al., 2011). Despite the inherent challenges, this step is of major importance to successfully identifying potential disease-causing or risk-generating variants (Dolled-Filhart et al., 2013). This is complicated by insertions or deletions (indels), which are especially problematic in the case of alignments performed by algorithms that disallow gaps, where there are significant PCR artefacts or variable GC content in short reads (in the case of single-end sequencing) and variable quality scores between bases; these generally degrading towards the end of individual reads (Nielsen et al., 2011, Dolled-Filhart et al., 2013).

The results of alignment and variant calling are thousands of differences between the reference genome and that under investigation (Nielsen et al., 2011). At this stage, these are filtered based on known variants in the population, and the results investigated for biological plausibility in leading to the potential disease pathology under investigation (Nielsen et al., 2011, Dolled-Filhart et al., 2013).

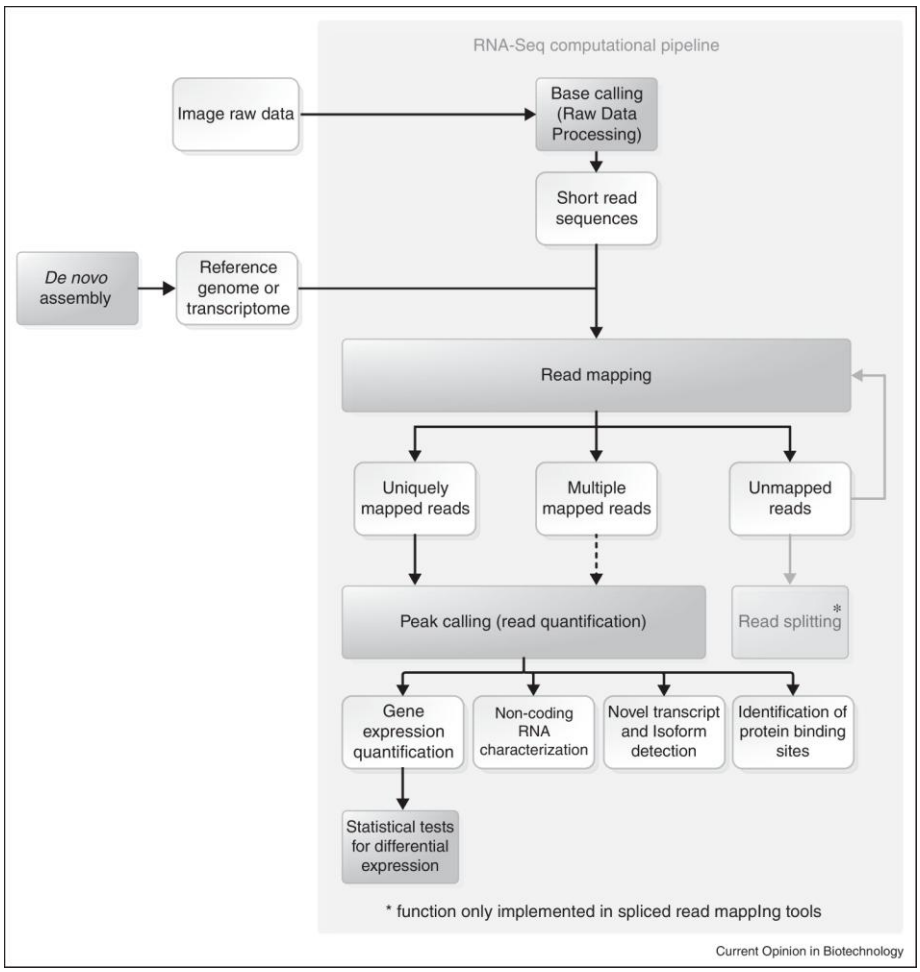


Figure 1.5. RNA-sequence analysis pipeline. (Figure from Mutz et al., 2013).

The use of NGS for generating a snapshot of the transcriptome is a valuable tool to facilitate gene annotation and identification of splicing variants; possessing greater sensitivity than microarrays and having increased facility over real-time PCR with lower costs (Mutz et al., 2013). As with DNA-seq, the process of RNA-seq investigation (Fig. 1.5) depends upon a well annotated reference transcriptome to underpin the analysis (Mutz et al., 2013).

1.3.4. Single Nucleotide Polymorphisms

Comparison of DNA sequences between individuals can reveal positions at which a specific, single base residue is variable (reviewed by LaFramboise, 2009). These single nucleotide polymorphisms (SNPs) are highly abundant, occurring approximately 1 in every 1000 bases of the human genome (Syvanen, 2001). Where these changes occur in a biologically active genetic region (e.g. coding regions of genes which generate a functional or structural protein) they can be sufficient to cause monogenic disorders, such as cystic fibrosis, Limb Girdle Muscular Dystrophy Type 2L (LGMD2L) or Duchenne Muscular Dystrophy (DMD) through interruption of normal gene expression

(Den Dunnen et al., 1989, Kerem et al., 1990, Hicks et al., 2011). This may act to directly cause disease, or may result in the increased risk of disease development, based on additional genetic and environmental factors.

In recent years, SNPs have been utilised to assist in identification of causative disease genes and to narrow genetic regions of interest, for which they are ideally suited due to their frequency and relatively uniform distribution throughout the mammalian genome (Sellick et al., 2004, Hardy and Singleton, 2009). They are well suited to be used as proxies for specific genomic regions due to the 'linkage' of nearby genomic regions to one another; especially those which have been validated as highly predictive of local genomic regions and are strongly conserved in a population (referred to as 'tag' SNPs) (The International HapMap Consortium, 2005, LaFramboise, 2009).

Analyses using SNPs represent both reduced cost and increased adaptability over microsatellite marker analysis (Sellick et al., 2004, Halperin et al., 2005, Zhang et al., 2005, Nielsen et al., 2011).

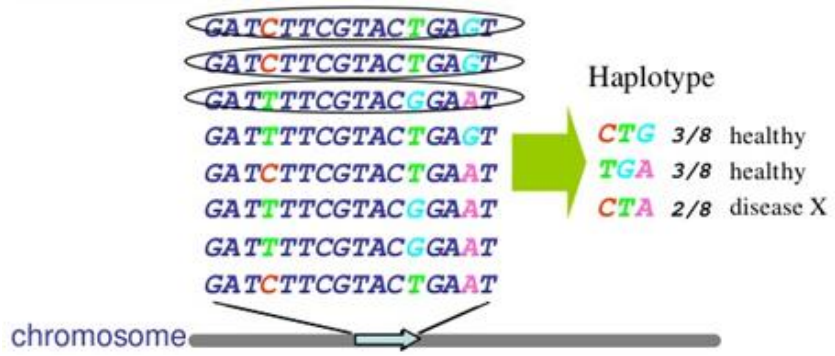


Figure 1.6. Haplotype as unit of statistical testing. Disease can be associated with a particular haplotype. (Figure from Chao, 2012).

Over a particular genomic region, the particular combination of alleles at these single nucleotide polymorphic sites is referred to as the haplotype (LaFramboise, 2009). For a given haplotype there may be an associated risk of disease or protective mechanism (Fig. 1.6), which may result not just from the influence of a particular SNP allele, but of the combination and potential interaction of multiple polymorphic nucleotides (LaFramboise, 2009).

The use of SNPs is now commonplace in human genomics, and the widespread adoption of SNPs as targets in genomic analysis has led to the development of genome-wide SNP arrays for several species (Fan et al., 2010). These have succeeded in identifying genetic polymorphisms either causative or risk-associated for a range of

human diseases, such as bipolar disorder, Crohn's disease and diabetes (Wellcome Trust Case Control Consortium, 2007).

1.3.5. Linkage

As an integral part of positional cloning, linkage analysis is the basis of identifying diseases which exhibit Mendelian inheritance patterns, in which a single mutation causing disease is propagated generationally (Botstein and Risch, 2003). Positional cloning identifies specific chromosomal regions which are transmitted within families, along with a disease phenotype of interest, thus demonstrating genetic linkage of a region with the goal of identification of the disease-causing mutation (Altmüller et al., 2001). It is thus a family-based approach limited to Mendelian disorders, in contrast with association analysis (further explained later) which does not require a defined pedigree and which can be applied to complex diseases (Altmüller et al., 2001).

It has been used to successfully identify the causative loci in a range of monogenic diseases, leading to potential diagnostics, treatments and possible cures, the most well-known of which being cystic fibrosis and Huntington's disease (Kerem et al., 1989, Di Maio et al., 1992)

Linkage is a measure of how connected individual loci are in the genome, based on their physical proximity (Cui et al., 2010). Those regions between which a recombination event is unlikely to occur can be considered to be 'linked' in that a given genetic feature will co-segregate with nearby genetic features during meiotic segregation (Cui et al., 2010). Thus we can use linked genetic markers like SNPs to predict the presence of a local genotype with a high degree of confidence (Sellick et al., 2004, Zhang et al., 2005). The measure of linkage is usually indicated by Log Of the Odds (LOD) scores, with greater than 3.0 being considered reasonable evidence for linkage, representing a 1000 to 1 chance of such a such a statistical result was arrived upon by chance alone (Teufel et al., 2006, Fukuda et al., 2009, Cui et al., 2010).

As a means to investigate genotypic variation underlying phenotypic variation linkage analysis is limited by several concerns; where misdiagnoses, heterogeneity, or complex inheritance are present a linkage-based analysis may well fail (Botstein and Risch, 2003).

1.3.6. Linkage Disequilibrium and Genetic Association Analysis

Linkage Disequilibrium (LD) is a statistical construct measuring the correlation of alleles at different loci (Visscher et al., 2012). It does not assume a necessary physical proximity, though in general loci that are closer together will possess higher LD; two genetic features separated by a large map distance may nevertheless possess tight correlation, meaning that the presence of a specific genotype in one region can be predictive of a corresponding genotype in another (Devlin and Risch, 1995, Cui et al., 2010). LD between loci is strengthened through the evolutionary forces of mutation, drift and selection and is weakened by recombination events (Hartl and Clark, 1997). The larger the effective population size, the weaker will be the LD for a given map distance (Hill and Robertson, 1968).

Genetic association testing examines the association of a genetic variant (e.g. SNP allele, microsatellite, or haplotype) with some phenotypic condition by statistical analysis, thus potentially providing new insights into related pathological pathways (Visscher et al., 2012). Where this correlation is able to be understood to be biologically relevant, it may lead to a greater understanding of disease pathology and the causative effects of the variant.

1.4. Neuromuscular Disorders

Neuromuscular disorders (NMD) are defined by involvement of both the peripheral nervous system and the muscles and they affect all age groups; from fetuses to the elderly (Emery, 1991, Laing, 2012). A 1991 worldwide review estimated that their prevalence amongst both sexes was approximately 1 in 3500 people (Emery, 1991). They are generally inherited and characterised by chronic, often progressive, loss of skeletal muscle strength leading to a loss of function, and are life-long diseases (Tews, 2002, Mercuri et al., 2007, Laing, 2012). In advanced cases, respiratory and cardiac failure ensues, resulting in early death (Rochester and Esau, 1994, Gozal, 2000, McDonald, 2002). For most neuromuscular disorders the treatment options are extremely limited and reduced to mitigation of symptoms rather than curative interventions. These include walking/standing frames, feeding tubes and adaptive technologies to help affected individuals carry out daily tasks, drug treatments and surgeries to lessen the worst of the resulting symptoms and usually require the attention of referred specialists (Muscular Dystrophy Association, 2013). In the case of autoimmune disorders drugs treatments can slow or halt the pathology, and prednisone

is currently prescribed for individuals with Duchenne's muscular dystrophy and some other conditions to slow muscle breakdown (Weiner, 2004, Beenakker et al., 2005, Muscular Dystrophy Association, 2013).

The term "NMD" encapsulates a wide range of conditions and involves a large variety of biological pathways; as of January 2013 more than 300 different genes have been shown to be causative of one or more of these conditions (http://www.musclegenetable.fr). This includes those in which the nerves and neuromuscular junctions are affected, such as in multiple sclerosis, those having neurological involvement such as in Parkinson's disease, or those that directly affect the skeletal muscle, such as in a range of Muscular Dystrophies (MDs) (Campbell, 1995, Emery, 2002, Lutton et al., 2003, Bertini et al., 2011, Pedrosa and Timmermann, 2013). The MDs can result in impaired cognition and neurological development, pathological cardiac involvement, hearing loss, eye malformation or degradation, loss of respiratory function, and central nervous system pathology (Dubowitz, 1965, Zellweger, 1965, Cohen et al., 1968, Kondo-lida et al., 1999, Gozal, 2000).

1.4.1. Muscular Dystrophy

MD normally presents with selective defects of the skeletal muscle proteins, along with muscle cell and tissue death (Emery, 2002). It is usually progressive, and muscular dystrophy is estimated to affect 1 in 3500 people (Emery, 1991). MDs are genetic in nature (over 30 different genes having been identified as causative) but are not always inherited; it is estimated that approximately one third of MDs are the result of a de novo mutation (Laing, 2012). The cost of these debilitating disorders in Australia is estimated to be approximately $1.5 billion per year, with a case burden three times greater than that of multiple sclerosis and ten times greater than diabetes; the disability adjusted life years per case is greater for these conditions than for any of National Health Priority Area (Access Economics, 2007).

1.4.2. Congenital Muscular Dystrophy

Congenital muscular dystrophies are a group of MDs that are usually present at birth and are genetic in nature, being therefore heritable (Bertini et al., 2011).

Following is a discussion of four MDs of special relevance to this project.

1.4.3. Duchenne Muscular Dystrophy

Duchenne Muscular Dystrophy (DMD) is a heritable disease caused by mutations in the X-linked dystrophin gene (*DMD*); this being the largest gene in the human genome (Kwiatkowska and Slomski, 1992, Blake et al., 2002). The disease has the second highest incidence of all inherited pathologies, affecting one in 3300 live male births (Emery, 1991). One in 10 000 haploid genomes will develop such a mutation *de novo*, meaning that genetic screening will never entirely eliminate the incidence of the disease; one third of all new cases are the result of such mutations (Barbujani et al., 1990). The lack of dystrophin protein leads to a breakdown of the necessary anchoring of sarcolemma protein to the underlying cytoskeleton in skeletal muscle, which results in increased damage to muscle cells and their breakdown (Fig. 1.7) with involved intellectual impairment (Zellweger, 1965, Karagan, 1979, Leibowitz and Dubowitz, 1981, Bresolin et al., 1994). This condition becomes noticeable from ~5 years of age and leads to a progressive loss of motor function, the majority of patients becoming wheelchair bound by the age of 12 and developing terminal cardiorespiratory complications by their second decade (Nowak and Davies, 2004).

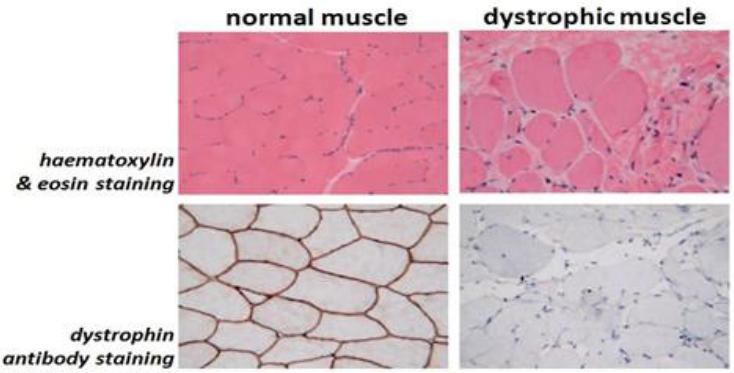


Figure 1.7. Histology of healthy and dystrophic muscle in Duchenne muscular dystrophy. Characteristic dystrophic muscle features are the result of successive rounds of degeneration and regeneration; including central nuclei, variation in fiber size and build-up of connective tissue between muscle fibres. Brown staining dystrophin is noticeably absent in dystrophic muscle tissue. (Figure adapted from Davies and Nowak, 2006).

1.4.4. Myotonic Dystrophy

Together types I and II myotonic dystrophy (MyD) affect one in 8000 people (Longman, 2006). The disease follows an autosomal dominant inheritance pattern and is a slowly progressive, multi-system disorder (Longman, 2006). It affects not only skeletal muscle but also cardiac and smooth muscle, the eyes, the endocrine and the nervous systems (Longman, 2006, Ashizawa and Sarkar, 2011).

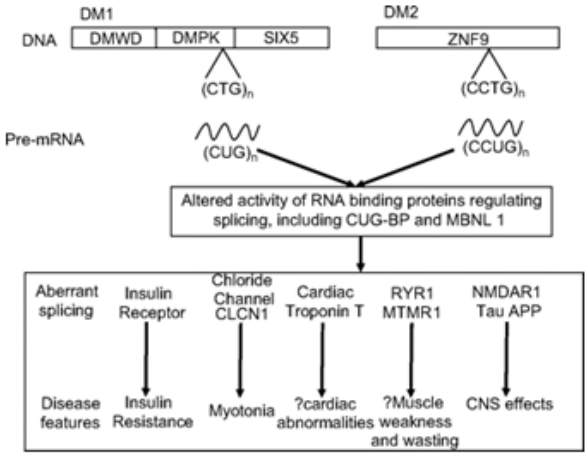


Figure 1.8. Genetics and pathology of myotonic dystrophy types I & II. (Figure from Turner and Hilton-Jones, 2010).

There is associated neurological pathology and cardiorespiratory complication (Turner and Hilton-Jones, 2010). The condition is the result of expanded repeats in either *DMPK* (type I) or *ZNF9* (type II) and severity is correlated with the size of the expanded repeat sequence (Fig. 1.8) (Brook et al., 1992, Klesert et al., 1997, Turner and Hilton-Jones, 2010, Ashizawa and Sarkar, 2011, Morales et al., 2012).

1.4.5. Nemaline Myopathy

The congenital myopathies like Nemaline Myopathy (NM) are defined based on structural abnormalities of the muscle fibres (North et al., 1997). NM can be inherited through either an autosomal dominant or recessive model and shows heterogeneity, with several causative mutations identified in genes generating components of the sarcomeric thin filaments, and, recently, in *KLH40*, the exact function of which is not yet known (Ilkovski et al., 2001, Ravenscroft et al., 2013). It presents with selective muscle weakness (Fig. 1.9) and nemaline bodies (Fig. 1.10) (Ilkovski et al., 2001).

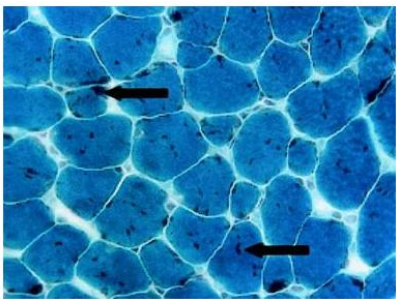


Figure 1.9. (left) Child affected with nemaline myopathy.

Figure 1.10. (right) Gomori trichrome staining in nemaline myopathy. Nemaline bodies indicated. (Adapted from Ilkovski et al., 2001).

1.4.6. Limb Girdle Muscular Dystrophy Type 2

Limb girdle MD (LGMD) is a heterogeneous collection of nearly 20 different diseases involving mutations at more than 50 genetic loci; depending on the subtype it has been shown to follow autosomal dominant (type 1) or autosomal recessive (type 2) inheritance (Nigro, 2003, Zatz et al., 2003). The disorder is rare; the incidence in the wider population for all LGMDs together being estimated at approximately 4-6 per 100 000 (van der Kooi et al., 1996, Urtasun et al., 1998).

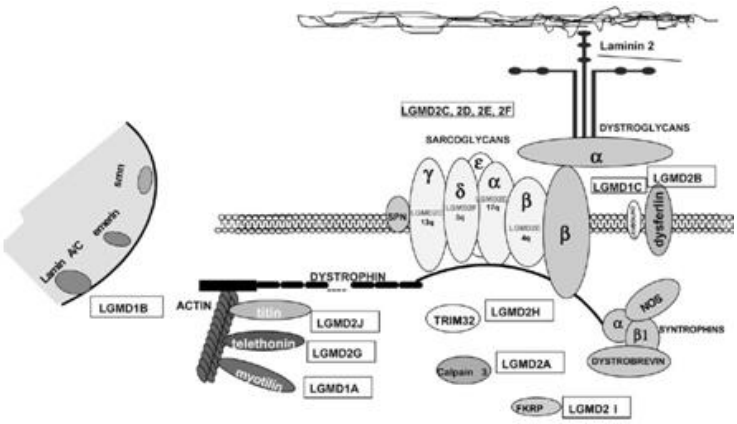


Figure 1.11. Proteins involved in pathogenesis of limb girdle muscular dystrophy. (Figure from Zatz et al., 2003).

Individuals with LGMD generally show muscle weakness and wasting restricted to the limb musculature (proximal greater than distal) and muscle degeneration/regeneration evident from tissue biopsy (Laval and Bushby, 2004). The molecular mechanisms involved in some subtypes are indicated (Fig. 1.11). Type 2G has been reported to be associated with nemaline bodies in affected skeletal muscle tissue (Paim et al., 2013). Onset, progression, weakness and wasting are highly variable between individuals and genetic subtypes (Laval and Bushby, 2004).

Limb Girdle Muscular Dystrophy type 2 (LGMD2) has been linked to mutations in 15 known genes (Table 1.1) and, depending on the subtype, symptoms show marked intrafamilial and interfamilial variability (van der Kooi et al., 1996). Possible treatments under investigation include exon skipping through gene therapy and cortisone therapy for types 2F and I (Danièle et al., 2007).

Table 1.1. Subtypes of limb girdle muscular dystrophy type 2, with associated genomic positions, affected genes and associated proteins (Adapted from Laval and Bushby, 2004)

LGMD2 Type	Chromosome	Gene	Protein
A	15q15.1-q21.1	*CAPN3*	calpain-3
B	2p13.3-p13.1	*DYSF*	dysferlin
C	13q12	*SGCG*	gamma-sarcoglycan
D	17q12-q21.3	*SGCA*	alpha-sarcoglycan
E	4q12	*SGCB*	beta-sarcoglycan
F	5q33	*SGCD*	delta-sarcoglycan
G	17q12	*TCAP*	telethonin
H	9q31-q34.1	*TRIM32*	tripartite motif protein (TMP-32)
I	19q13.3	*FKRP*	fukutin-related protein
J	2q24.3	*TTN*	titin
K	9q34.1	*POMT1*	protein O-mannosyltransferase
L	9q31	*FKTN*	fukutin
M	1p34-p33	*POMGNT1*	protein O-linked-mannose beta-1,2-N-acetylglucosaminyltransferase 1
N	14q24.3	*PMT2*	protein O-mannosyltransferase 2

1.5. A Sheep Model of Muscular Dystrophy

1.5.1. The OCPMD Flock

A naturally occurring animal model of a congenital muscular dystrophy was discovered in south-western Australia in the 1950's (McGavin and Baynes, 1969) and has been preserved as a research colony in the intervening years; bred and pedigree tracked by researchers and staff at both Murdoch University and the University of Western Australia. The Ovine Congenital Progressive Muscular Dystrophy (OCPMD) flock provides a singular opportunity to investigate this new disease.

1.5.2. Inheritance

This disease follows an autosomal homozygous recessive inheritance pattern and the specific phenotype has not been reported to manifest in any example other than this pedigree (Richards et al., 1988a; personal communication with Prof. Nigel Laing). Affected sheep have been identified over four Australian states; initially in Queensland, and subsequently in Western Australia, New South Wales and Victoria (Dent et al., 1979, Richards et al., 1988b).

1.5.3. Disease Morphology

Characterised by abnormal myofiber morphology, progressive myofiber loss and adipose replacement (Fig. 1.12), affected animals present with congenital muscle weakness, and selective skeletal muscle tissue degrades over time (Richards et al., 1986, Richards et al., 1988a). The progressive fibrous and fatty tissue replacement in the

muscle fibers of the hip, stifle and hock results in an inability to effectively mobilise the hindquarter muscles, locking these joints in extension, leading to a stiffened gait (Richards et al., 1986). The disease preferentially affects type I muscle fibers, sharing this feature in common with myotonic dystrophy and certain congenital myopathies of humans (Richards et al., 1988a, North et al., 1997, Longman, 2006). Type I fibers are those which are 'slow-twitch'; they provide endurance and fatigue slowly, while type II fibers are 'fast-twitch'; function in short bursts of action and fatigue quickly (Cunningham, 2007). The affected fibers are proportionately higher in those muscles which affect posture, meaning that this is particularly affected in the pathology (Richards et al., 1986, Richards et al., 1988a).

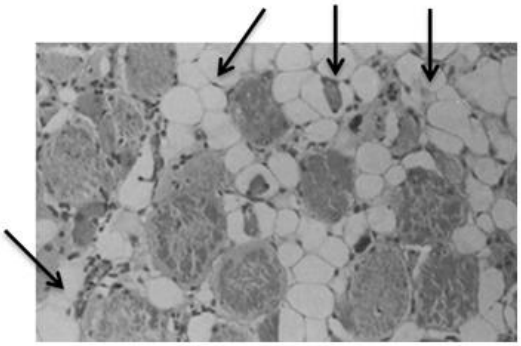


Figure 1.12. Adipose replacement in OCPMD-affected skeletal muscle tissue of the anconaeus of a 2 year old ewe from the ovine congenital muscular dystrophy flock, with the remaining fibers being severely dystrophic. Arrows indicate adipose cells. (Figure adapted from Richards et al., 1988a).

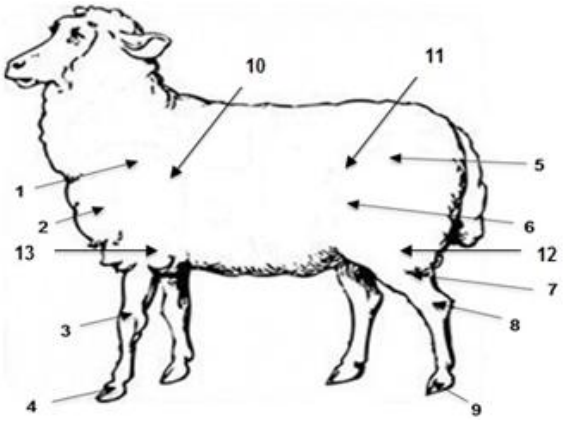


Figure 1.13. Most severely affected muscles in OCPMD-affected sheep. Labeled muscles as follows: **(1)** Flexors of shoulder **(2)** Flexors of carpus **(3)** Flexors of digits **(4)** Digits **(5)** Extensors of hip **(6)** Extensors of stifle **(7)** Extensors of hock **(8)** Flexors of digits **(9)** Digits **(10)** Triceps brachii **(11)** Vastus intermedius **(12)** Soleus **(13)** Anconaeus

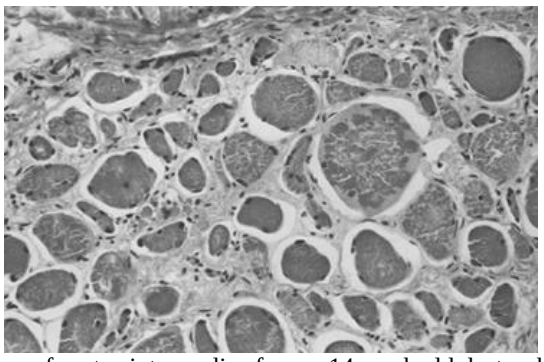


Figure 1.14. Transverse section of vastus intermedius from a 14-week old dystrophic ram showing the broad range of muscle fibers, rounded profiles, internal nuclei and peripheral sarcoplasmic masses typical of dystrophic tissue (Richards et al., 1988a).

The primary muscles affected include the extensors of the hip, stifle and hock joints, flexors and digits of the hind limb, the extensors of the elbow, flexors of the shoulder, carpus and digits of the fore limb (Richards et al., 1988a). Severe and consistent affection has been reported in a number of muscles (Fig. 1.13) (Richards et al., 1988a). The lesions described are consistent with the inherited muscular dystrophies, with dystrophic fibers usually distributed throughout the affected muscles, however being occasionally seen grouped into a focal loss of myofibrils, resulting in irregular sarcoplasmic masses (Fig. 1.14) (Richards et al., 1988a, North et al., 1997).

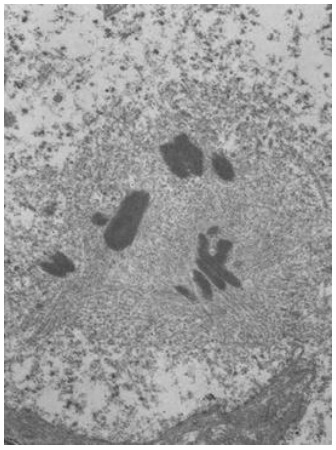


Figure 1.15. Disorganised fibrillar material containing small nemaline bodies beneath the cell membrane of dystrophic fiber (Richards and Passmore, 1989).

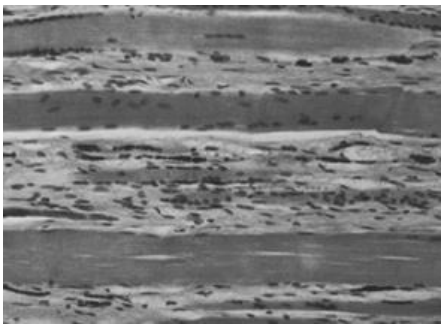


Figure 1.16. Longitudinal section of the vastus intermedius in an OCPMD affected sheep. Severe atrophy of fibers results in chains and clusters of small, dark-staining nuclei. (Richards et al., 1988a).

Respiratory infections stemming from decreased mucocilliary clearance and ruminal tympany as a result of affected skeletal muscle tissue in the diaphragm have also been observed (Richards et al., 1986).

Similar to NM, nemaline bodies are present in affected tissues (Fig. 1.15) (North et al., 1997). In otherwise intact muscle fibers, there are long rows of centrally placed muscle nuclei (Fig. 1.16) as in myotonic dystrophy (McGavin and Baynes, 1969).

There is significant variation in the severity and rate of progression between affected individuals (Richards et al., 1986).

1.5.4. Diagnosis

Other than muscle pathology and muscle weakness, markers for disease include elevated levels of creatine phosphokinase (CPK) subsequent to exercise in affected animals as compared to controls, as well as increased resting CPK, lactic dehydrogenase (LDH) and sorbitol dehydrogenase (SDH) concentrations (though CPK levels are not significantly elevated until onset of clear clinical disease symptoms) (Dent et al., 1979, Richards et al., 1986).

1.5.5. Gene Investigations to Date

Attempts have been made to characterise the OCPMD disease model using traditional methods of investigation. These include histology and muscle biopsy, microsatellite marker analysis and western blotting to examine a fairly extensive panel of the known MD genes (personal communication from A/Prof. Kristen Nowak).

1.6. Comparisons of the Ovine Model with Human Disease

1.6.1. Commonalities with Human Disease

The distribution of lesions in OCPMD is consistent with the inherited MDs, and there are several similarities to specific MD diseases. Histopathology was noted by Richards et al. (1986) to share marked similarity to that of MyD, and the presence of muscle weakness with nemaline bodies in affected skeletal muscle tissue is similar to that observed in NM (North et al., 1997). There are notable similarities to some subtypes of LGMD2, and the presence of nemaline bodies was recently reported in LGMD2G (Paim et al., 2013). In addition, the absence of cardiac and smooth muscle involvement, adipose replacement in affected tissue, diaphragm skeletal muscle involvement, and autosomal inheritance pattern with congenital onset is consistent with that of NM (North et al., 1997). The congenital nature of disease presentation invites comparison with several of the congenital MDs and NM, but this collection of pathological symptoms has not previously been observed in a single human disease (personal communication from A/Prof. Kristen Nowak).

1.6.2. Differences to Human Disease

Despite the histopathological commonalities between myotonic dystrophy (MyD) and OCPMD, other hallmarks of this human disease are absent in the pathology. MyD is associated with the persistence of action potential activity in affected muscles after cessation of stimulus, has a variable age of onset (between infantile up to the fourth decade) and with a high incidence of lenticular cataracts, myocardial lesions and testicular atrophy, all of which are absent in the OCPMD disease (Kakulas, 1985).

The differences between NM and OCPMD are less pronounced, yet dystrophic fibers are not usually a feature of NM, and the known causative genes for this disease have been largely eliminated from consideration through traditional gene investigations (personal communication from A/Prof. Kristen Nowak and Prof. Nigel Laing).

1.7. Bioinformatics

Bioinformatics represents the nexus of information technology and the biological sciences. Recent rapid advances in computing power, coupled with similar advances in biotechnologies, has led to vast amounts of data being generated routinely for scientific (and other) studies. Traditional computing software and hardware tools have been outstripped by data production and complexity (Luscombe et al., 2001, Magi et al., 2010). As a result, there is an increasing need for specialised software tools and platforms on which to carry out computationally-intensive analyses. These include such approaches as using Genome-Wide Association Studies (GWAS), Next-Generation Sequencing (NGS) analysis and processing, and genomic database handling, each described previously (Luscombe et al., 2001, Teufel et al., 2006, Magi et al., 2010, Moore et al., 2010, Dolled-Filhart et al., 2013). Along with these tools comes the need for specialised scientists with both an understanding of the underlying physiology and of the technology required to tease out meaningful information from the whole, giving rise to the critical importance of bioinformatics researchers and resources.

1.7.1. Origins of Bioinformatics

The field of computational biology is the precursor to bioinformatics (Hagen, 2000). While its major growth came out of the human genome project and associated investigations into additional genomes, computers were an integral part of biological investigation for a full decade before DNA sequencing became feasible (Boguski, 1998)

During the 1960's, the expanding collection of amino acid sequences developed by the biological research community represented interesting problems which were best investigated using computationally-based approaches (reviewed by Hagen, 2000). High-speed digital computing from weapons-research programs were becoming widely available and thus became an effective tool in understanding complex biological questions such as those posed by protein biochemists (Hagen, 2000). The development of the FORTRAN programs to determine the sequence of protein molecules and the development of algorithms to understand sequence homology and alignment accounting for deletions and insertions in the early 1960s represented watershed moments in bioinformatics development, though the field had not yet been so named (Dayhoff, 1962, Dayhoff, 1965, Fitch, 1966). By 1970, computational biologists had developed all the major precursors to techniques for investigating nucleic acids, though these were all originally created to study proteins (reviewed by Hagen, 2000).

The impact the genome projects was especially felt in the growth of databanks for molecular biological information such as GenBank, EMBL, DDBJ, PR and SWISS-PROT (Kanehisa and Bork, 2003). Both the diverse, large-scale data and associated need for sophisticated methods of handling and interrogation, and the changing landscape of informatics tools such as the internet and computing technology synergistically drove bioinformatics development in the following years (reviewed by Hagen, 2000). The introduction of sequence database tools such as FASTA in the early 1980s and BLAST in 1985 were landmark events in the elucidation of useful information from massive data generation (Lipman and Pearson, 1985, Altschul et al., 1990).

The years since have seen bioinformatics grow at a pace far outstripping Moore's law, as the data generation of technologies has similarly advanced, with many databases doubling in size every 15 months (Benson et al., 2000).

1.7.2. Importance of Bioinformatics to this Project

For the required work of this project bioinformatics represents a highly necessary component. NGS data is far too large to utilise without such specialised tools and, when dealing with data from the SNP genotyping of a number of individuals, being able to effectively manipulate the data for interrogation is a non-trivial step.

Carrying out alignments and variant calling from NGS data requires data manipulation in terms of changing file format and layouts so that specific tools can be applied for each discrete step in the analyses, and being able to visualise the results of these alignments and variants through the use of tools designed to do so requires a degree of expertise. For these analyses the data size and complexity is too great for simple approaches like manually editing a text file by hand, or opening in a spreadsheet tool.

Linkage and association analysis is computationally intensive, and requires tools capable of conducting tests with precise control of variables, in addition to a computational framework on which to run the analyses.

By leveraging the power of program scripts, specialised bioinformatics tools based in Unix, and the large variety of online bioinformatics resources, we were able to elucidate useful information from the datasets we were provided.

1.8. Research hypothesis and aims

The hypothesis of my research project was that the cause of disease in the OCPMD flock is genetic, and that the causative gene/s could be identified through bioinformatics analyses and further investigated through molecular biological applications.

This was intended to be accomplished by fulfilling the following aims:

i.) To utilise the SNP genotype data on informative members of the OCPMD flock to undertake combined linkage analysis and association mapping and thus identify strong statistical candidates to locate a causative disease gene or genes.

ii.) To investigate these gene and SNP candidates for the most likely candidate genes based on plausible biological explanations derived from the existing literature and database information.

iii.) To confirm the *in silico* findings through homozygosity mapping of the affected, carrier and control Illumina whole-genome sequencing, in combination with the results of the preliminary GWAS from our collaborator at the CSIRO.

iv.) Assuming the identification of a strong statistical and plausible biological candidate gene/s and time permitting, to investigate the disease pathology and implicated pathways using molecular biological techniques.

1.9. Significance of the Project

The MD sheep model provides a unique opportunity to study the various associated pathologies, enabling much more opportunity for skeletal muscle biopsy and the resulting pathway analysis than has been previously possible in smaller mammalian models of MD. To date, the selective nature of MD pathogenesis in humans has not been well understood, and the development of this model to assist investigation is of major importance to the field.

It also provides a unique opportunity to trial potential therapies in an animal model much more representative of human skeletal muscle mass than any currently available. With the recent advances in the sheep reference genome, we are now much more likely to be able to pinpoint the specific genetic involvement of this particular disease, providing the potential to improve our understanding of MD pathogenesis in general, in addition to characterizing this particular model.

The use of a bioinformatics-based methodology provides us with the ability to use the existing pedigree, an improved virtual genome, and recently developed software infrastructure to discover the genetic basis of this unique disease presentation, and extrapolate this to human disease and treatment research.

2. Materials and Methods

This chapter first details an overview of the OCPMD flock along with the genetics and genomics considerations for the project. It then explains the bioinformatics workflow for investigation and details the various analyses employed to interrogate the genetic data available. The chapter concludes with the methodology for investigation of the best candidate gene using molecular biological techniques.

2.1. The Ovine Congenital Progressive Muscular Dystrophy Flock

2.1.1. History

The first clinical cases were seen in 1953, but it wasn't until 1958 that a diagnosis formed from the histological evidence was made (McGavin and Baynes, 1969). Individuals known to be affected by the disorder were bred to produce the Ovine Congenital Progressive Muscular Dystrophy (OCPMD) flock and their breeding carefully tracked and directed over decades in order to sustain an informative flock. During this time the disease phenotype, including pathology, was investigated through histological examination and clinical observation of affected individuals (Richards et al., 1986, Richards et al., 1988a, Richards and Passmore, 1989). The sheep were subsequently located to Murdoch University in ~ 1990 and cared for there, before being moved to the UWA Shenton Park Sheep Research Facility in 2012.

2.1.2. Description and definition of phenotype

The disease was reported as a congenital muscular dystrophy in 1969 (McGavin and Baynes, 1969) and the clinical presentation further elucidated in several papers in the late 1980's (Richards et al., 1986, Richards et al., 1988a, Richards and Passmore, 1989).

OCPMD bears similarities in presentation to the human diseases nemaline myopathy and myotonic dystrophy, but the specific combination of pathologies has not previously been observed in a single disease in humans.

OCPMD was reported to follow an autosomal recessive inheritance (Richards et al., 1988b). While there has been some variability reported in the severity of affection between individuals of the flock, a recessive inheritance with no penetrance effects in carriers was chosen as the best fitting analysis model. Recent evidence has suggested some mild pathology in certain skeletal muscles of two aged individuals (one carrier and one 'unaffected' of unknown genetic status), which was not suggested clinically

prior to necropsy (Report by veterinary pathologist Dr Amanda O'Hara). Additional necropsies in age-matched controls may be necessary to determine whether 'mild' affection in 'carriers' exists (also known as 'manifesting carrier') or if the pathologies exhibited were the result of ageing.

Being a recessively inherited disease, individuals can be classified as being:

(a) Affected (b) Carrier (c) Normal or (d) Unknown.

(a) 'Affected' phenotype

Defined by fulfilment of all of the following criteria, although sometimes not all criteria are able to be scored whilst the sheep is alive:

(i) Noticeably stiffened gait in the hind limbs, leading to a reduced ability to run effectively in comparison to unaffected individuals
(ii) Skeletal muscle tissue wasting in affected muscle groups with adipose replacement
(iii) Presence of nemaline bodies in affected muscle tissues at biopsy
(iv) Raised creatine kinase at rest, and particularly after exercise in comparison with unaffected individuals

'Affected' individuals can be:

(i) Descendent from two known affected individuals (100% chance).
(ii) Descendent from an affected and carrier individual, along with phenotypic presentation of the disease pathology (50% chance)
(iii) Descendent from two known carrier individuals along with phenotypic presentation of the disease pathology (25% chance)

(b) 'Carrier' individuals do not show any clinical signs, and could be:
(i) Descendent from an affected and a carrier individual (50% chance),
(ii) Descendent from an affected and an unaffected individual (100% chance).

(c) 'Normal' individuals do not show any clinical signs, and could be:
(i) Descendent from two animals outside the directed pedigree (essentially 100% chance)
(ii) Descendent from two carrier individuals (25% chance)

(d) 'Unknown' individuals are those that genetically could be a carrier or normal. They do not show any overt signs of being affected. They might be:

(i) Descendent from two carrier individuals (25% chance to be a carrier, 25% chance to be normal)
(ii) Descent from a carrier and a normal (50% chance to be a carrier, 50% chance to be normal)

2.1.3. Overview of Pedigree

While the OCPMD flock has been cared for and examined for many decades, there are data missing in regards to some parentage, lines of inheritance and disease state.

The pedigree (Fig. 2.1.) has followed a careful pattern of directed breeding to ensure that new individuals used to continue the line of inheritance are of a known disease state in order to enable assumptions about the genotypes of these individuals in relation to the putative disease gene.

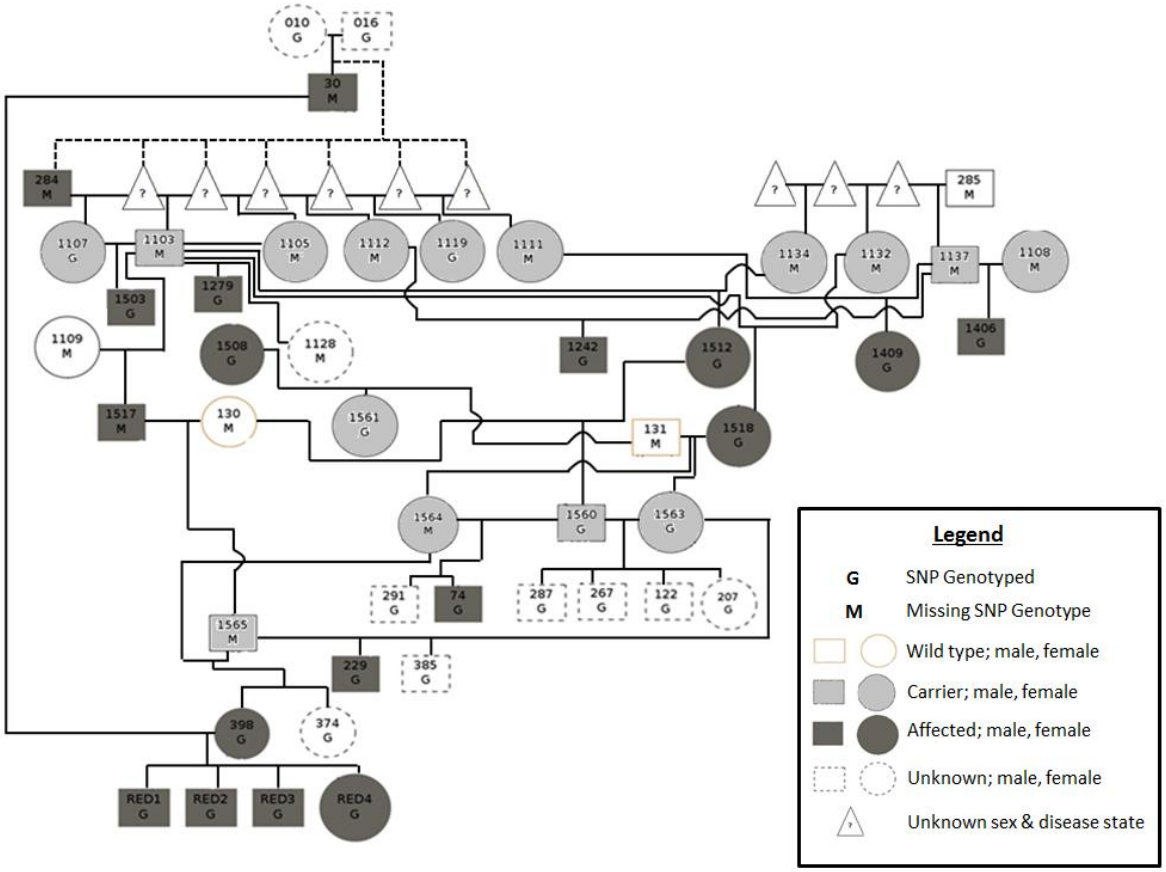


Figure 2.1. Full OCPMD pedigree inclusive of all information.

A small number of individuals have had semen samples collected and some have been used to artificially inseminate eggs harvested from ewes in the flock to produce the last generation of the pedigree. There are additional fertilised frozen eggs available for subsequent implantation into a ewe.

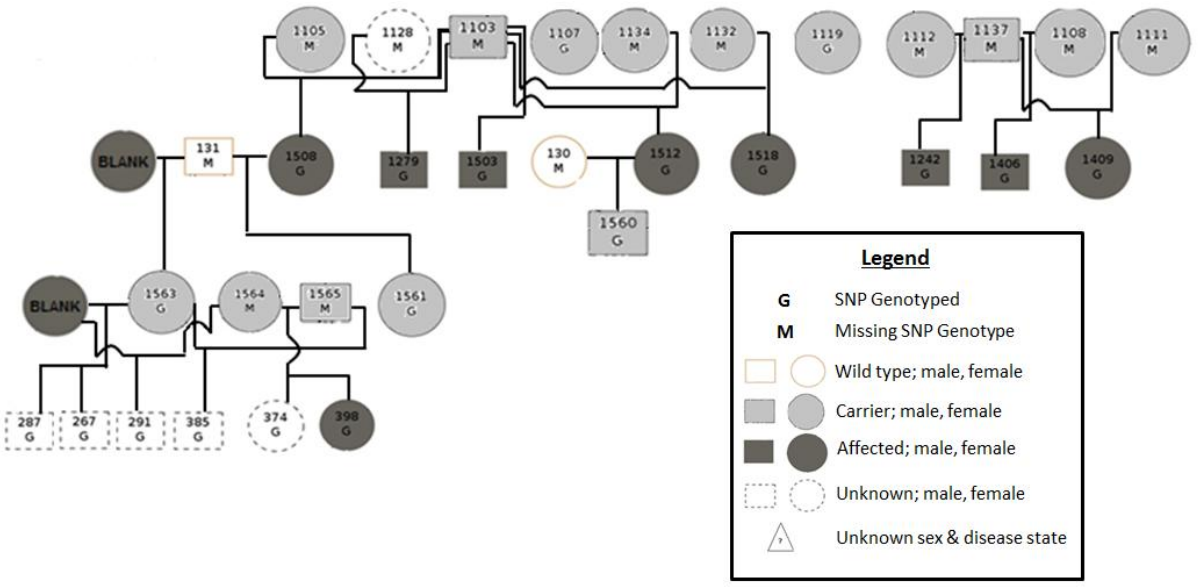


Figure 2.2. Pedigree of initial association analysis and homozygosity mapping.

Some reported inheritance was demonstrated to result in impossible recombinations between genotypes by preliminary linkage analysis, using the software packages Pedstats and Merlin (Abecasis et al., 2002, Wigginton and Abecasis, 2005). This may have resulted from some samples being mislabelled when sent for genotyping, or previously to that point (e.g. lambs attributed to a certain mother may not have been; a mix up when the blood samples were taken). These inheritance errors necessitated using some 'blank' individuals for the pedigrees utilised in the initial analysis (Fig. 2.2.) and that of the final analysis (Fig. 2.3.)

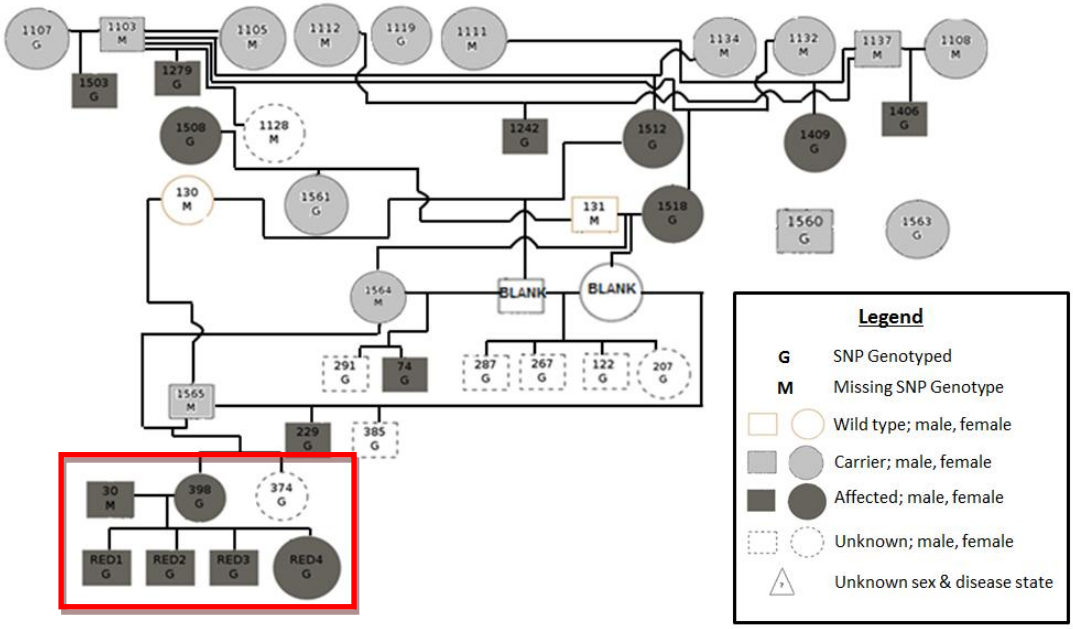


Figure 2.3. Pedigree of final association analysis and homozygosity mapping. Outlined individuals are the nuclear family used for linkage analysis.

Several additional individuals were genotyped on the SNP array in order to provide additional strength for the association and linkage analyses. This included four sib-pairs (Designated 'Red 1 – Red 4') in which one parent was of known genotype and affection status (the other parent was a known affected, but his DNA was not available for genotyping).

2.1.4. Collection of biological samples

The University of Western Australia Ethics Committee approved the maintenance of the flock, and all procedures performed on members of the flock associated with this project. Blood samples were collected from all pedigree members during the maintenance of the flock in order to enable genotyping. Semen samples were collected from certain males in recent generations to facilitate directed breeding through artificial fertilisation at later dates, and in particular to ensure preservation of the flock.

Samples of skeletal muscle tissue from two affected sheep (#398 and #74), one known carrier (#1563) and one clinically unaffected sheep with unknown genotype (#229) were collected at post-mortem and preserved by snap-freezing in liquid nitrogen and stored at -80 ^{o}C.

2.1.5. Histology and biomarkers

The gold standard method for determining the affection status of any individuals from this flock has been biopsy of soleus skeletal muscle tissue to confirm pathology, as this muscle shows consistent morphological changes in all affected individuals (Richards et al., 1986). Additionally it is a muscle that can be excised during anaesthesia and not cause any detrimental effect on the sheep afterwards. Affected soleus muscle is characterised by skeletal muscle wasting and progressive fibrous and fatty tissue replacement, and the presence of nemaline bodies under histological examination.

Creatine kinase was generally elevated for affected individuals beyond that of unaffected individuals at rest or after exercise (Richards et al., 1986). However, the levels of this enzyme are not considered a reliable enough measure of disease status due to a reduction in the degree of creatine kinase elevation as the disease progresses (e.g. an exhaustion of the skeletal muscle that is degenerating, and therefore less release of creatine kinase) (personal communication with Dr. Amanda O'Hara).

2.2. Genetics and Genomics

2.2.1. Sheep Genome Project

The sheep genome project is ongoing and not finalised, however the latest build released for research purposes is version 3.1. (Archibald et al., 2010). There is no SNP database available such as dbSNP, and the annotation is not widely available for interrogation at this time, however it is available for direct download (http://www.livestockgenomics.csiro.au/sheep/oar3.1.php). It has been utilised in this project to assist investigation into genes implicated by linkage, association analysis and homozygosity mapping. Reported genes are based on predicted protein coding sequences.

The sheep reference genome version 3.1. was provided by Dr. James Kijas in FASTA format.

2.2.2. SNP Array

As part of the international sheep genome consortium's (ISGC) sheep reference genome work, an array containing 49 034 high quality SNPs for use in the investigation of sheep genetics was developed (http://www.sheephapmap.org/images/pag09_dalrymple.pdf). This was used in this project and will be referred to as 'the SNP array'.

Genotyping of samples on the SNP Array was performed by Dr. James Kijas' laboratory team and the results provided to this project. The array has 53 903 individual SNPs, for which 49 034 passed the laboratory quality control (QC). For each genotyped individual a file containing the genotyping results in PLINK format (.ped) and a file containing the genomic map position of each SNP (.map) was provided. A file containing the full collection of SNPs used for genotyping (.dat) was generated from the provided files.

The positions of the included SNPs were provided in the SNP genomic annotation file (http://www.livestockgenomics.csiro.au/sheep/oarv3/Oarv3.1.50kSNP.position.gff3). This included the positions at previous reference genome builds as well as the v. 3.1. positions.

Initial Analysis

SNP genotypes from 9 informative affected and 5 informative carrier individual SNP genotypes were made available for analysis. In addition, the genotypes of several individuals classified ‘unknown’ were provided (Fig. 2.2.).

Final Analysis

After genotyping of additional members of the flock was performed, SNP genotypes from 14 informative affected and 6 informative carrier individual genotypes were available for analysis, including those available in the initial analysis. In addition, several genotypes for ‘unknown’ individuals were provided (Fig. 2.3.).

2.2.3. Whole-Genome Sequencing

Two individuals from the OCPMD flock were whole-genome sequenced as part of a larger project conducted on 50 sheep by the ISGC on the Illumina platform and the results provided to this project. These sequences were aligned to the sheep reference version 3.0. using SAMtools and mpileup and provided in .bam format.

The binary files for these aligned sequences, being one affected individual (#398) and one carrier individual (#1560) were provided to this project along with one ‘normal’ merino sheep aligned to the same reference. These were of approximately 8x coverage.

2.3. Bioinformatics

Bioinformatics analysis required manipulation of data resources into the appropriate format for software utilisation. This process also involved a large degree of quality control and the adjustment of research strategies based of the resulting information. The workflow of analysis and candidate selection strategy was adjusted based on the quality of and limitations on the provided data.

2.3.1. Analysis

The goal of the bioinformatics analysis workflow (Fig. 2.4.) was to cast a wide net at each phase of the analysis (homozygosity mapping, association analysis, and linkage analysis) in order to develop a large candidate geneset for the putative disease mutation. All of the ‘best candidates’ from each approach were included in the geneset. The results of each of these approaches were then cross-referenced to find candidates for which multiple lines of evidence supported further examination, and those which were highly significant under one approach or moderately significant under multiple

approaches investigated by interrogation of published literature and publically available databases for a biologically plausible involvement in the OCPMD pathology.

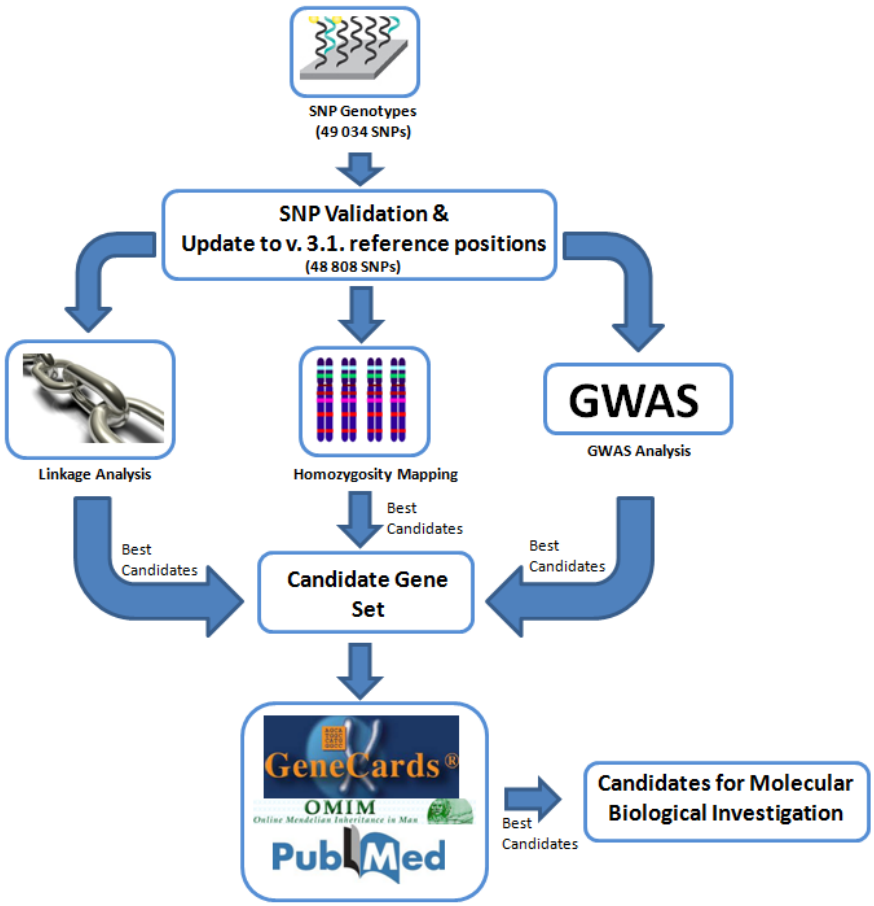


Figure 2.4. Flowchart for bioinformatics analyses leading to candidates for further investigation.

2.3.1.1. Homozygosity Mapping

Homozygosity mapping was carried out using a custom tool, written in perl: the 'Homozygosity_Mapper' (see appendix V for source code). While there are existing homozygosity mapping tools available, many of the tools are designed for human data only (e.g. fixing the expected number of chromosomes) and do not handle complex pedigrees, such as our OCPMD data. Thus we determined that it would be best to construct out own tool to perform this task.

The homozygosity mapper interrogated the dataset of SNP genotypes for consecutive sequences of homozygous SNPs, ignoring missing values, and reported all consecutive runs over an input threshold value for those of the specified subset. This included the genomic positions of SNPs. By specifying the included individuals by disease state I was able to discover regions of homozygosity specific to affected individuals within the pedigree. By including a 'wobble factor' into the tool I was able

to examine potential runs of homozygosity with an input number of 'mismatches' (SNPs which were homozygous in all but 1 or 2 individuals, for example – depending on the wobble value). This loosened the specificity, allowing for a single (or more) genotype error or mismatch, while identifying potential long stretches of homozygous markers which would otherwise not be observed.

The mapper then accessed the v. 3.1. sheep reference genome annotation file and used the reported SNP and gene positions to automate the reporting of which genes were included in these regions of homozygosity.

Initial Analysis

The initial analysis included 9 affected individuals to identify homozygous runs in all affected, and 5 carriers to determine if these homozygous regions were shared between both the affected and carrier groups (Fig. 2.2.).

The longest run of homozygosity observed in the dataset was with the following command:

```
$ perl Homozygosity_Mapper.pl
InitialPedigreeWithBlanks.ped
SNP50_3.1.ReferencePositions.map 48808SNPs.dat 2 0 8
```

This specified that only affected individuals (2) were to be included in the analysis, no wobble factor (0), and the reporting of sequential homozygous SNPs in a run of ≥8 (8).

The next longest runs of homozygosity in the dataset were observed with the following command:

```
$ perl Homozygosity_Mapper.pl
InitialPedigreeWithBlanks.ped
SNP50_3.1.ReferencePositions.map 48808SNPs.dat 2 0 5
```

This had the same parameters as the previous search, but reported all runs of homozygous SNPs ≥5.

To discover whether these regions of homozygosity were distinct to the affected individuals commands of the same parameters but specifying carriers only were also input:

```
$ perl Homozygosity_Mapper.pl
InitialPedigreeWithBlanks.ped
SNP50_3.1.ReferencePositions.map 48808SNPs.dat 1 0 5
```

The results of these analyses were then compared. Regions of homozygosity shared between carriers and affected were still considered for candidate gene selection, but less weight given to the genes within the shared regions.

Final Analysis

The final analysis included 14 affected individuals to identify homozygous runs specific to the affected and 6 carriers to determine if these homozygous regions were shared between both the affected and carrier groups (Fig. 2.3.).
The longest run of homozygosity was observed with the following command:

```
$ perl Homozygosity_Mapper.pl FinalPedigree.ped
SNP50_3.1.ReferencePositions.map 48808SNPs.dat 2 0 6
```

The next longest runs of homozygosity were observed with the following command:

```
   $ perl Homozygosity_Mapper.pl FinalPedigree.ped
SNP50_3.1.ReferencePositions.map 48808SNPs.dat 2 0 5
```

As with the initial analysis, the same commands were input specifying carrier individuals and the results compared with that of the affected to check whether the regions of homozygosity were shared by all the flock, or just the affected individuals.

The genes present in regions of SNP homozygosity specific to the affected individuals at the top two thresholds were included in the candidate geneset for further investigation. Homozygous runs which were only present in the carriers were discounted, while those shared by both the carriers and affected sheep were included in the candidate geneset, but given lesser weight than those limited to the affected individuals.

2.3.1.2. Association using PLINK

For the association analysis, we chose to use the PLINK software package (Purcell et al., 2007). PLINK is very widely used for GWAS analysis, and is capable of

performing family-based association analysis and handling the extended pedigrees from the OCPMD flock.

Initial Analysis

The initial association analysis using PLINK was limited by a low sample size and the exclusion of individuals that had an unclear phenotype or unknown genotype as related to the putative disease gene.

As already described, of those individuals that were genotyped and of known phenotype 9 were affected and 5 were carriers.

The recessive model was included in the analysis in accordance with the reported inheritance pattern for this disease.

Despite being part of the same large pedigree, in most cases SNP genotyped individuals were not connected into nuclear family units; to enable computational processing the 'discrete families' option was used. This broke the broader pedigree into small family units, but the analysis was essentially that of disconnected individuals.

As the map file for this analysis consisted of only 3 columns, the option of map3 was specified.

The included option within PLINK to specify sheep as the species was employed.

The data did not meet the assumption of minimum cell count number (≤ 5) for the χ^2 test in regard to the control (carrier) individuals. For this reason the option of reducing the minimum count in a cell to 0 was used.

Due to the small sample size, relatively sparse coverage of the genome, and the inability to use a test assuming the relatedness of these individuals, control for multiple testing was not able to be employed. This limited the investigation to being exploratory. Both adaptive and discrete permutation testing were attempted, but resulted in no single SNP reaching an adaptive significance level smaller than a p-value of 0.3. For this reason the control for multiple testing errors was created by the complementary bioinformatics analyses being mutually supportive of candidate gene selection, rather than any one approach being specifically controlled for type 1 errors.

Command:

```
  $ plink --ped InitialPedigreeWithBlanks.ped --map3 --
sheep --model --model-rec --dfam --cell 0 –map
SNP50_3.1.ReferencePositions.map
```

Final Analysis

While the final association analysis had greater numbers than the initial, the sample size was still a lower number than would have been ideal, and the exclusion of individuals who had unclear phenotype or unknown genotype as related to the putative disease gene further decreased the number of individuals for testing. This analysis included the members of the initial analysis.

Of those individuals which were genotyped and of known phenotype 14 were affected and 6 were carriers.

The recessive model was again included in analysis in accordance with the reported inheritance pattern for this disease.

As before, the discrete families option, sheep option, and reduction in minimum cell size to 0 was used. The analysis was run with the following command:

```
$ plink --ped FinalPedigree.ped --sheep --model --model-
rec --dfam --cell 0 --map
SNP50_3.1.ReferencePositions.map
```

For both the initial and final dataset, the results of the association analysis were extracted and those meeting a p-value significance level of ≤0.05 had their chromosomal base position and gene within which they occurred (or closest gene and distance to gene) inserted into the results. This was carried out by use of a custom tool written in perl which accessed the genome annotation file and SNP position file, called the ‘Plink_Parser’ (see appendix VII for source code).

Any SNP which did not occur within a gene or within a close distance was excluded from further analysis, as were any SNPs which occurred on the X chromosome, as the disease is autosomally inherited. SNPs which were found significant in the results of the initial dataset but not that of the final dataset, or were found to be less significant in the final dataset than initially, were excluded from the results. While the two analyses did

not satisfy the assumptions of independent testing, the time limitations of the project necessitated using the initial analysis to identify gene candidates prior to additional SNP array results being available. Thus, the initial and final analyses were used to complement one another. Those which were only slightly more significant in the final analysis than the initial results were excluded, in favour of those for which the p-value was decreased more substantially (~90% of the SNPs found to be more significant were of a p-value 50% smaller than in the initial dataset).

Two copies of these results were then made; one sorted by p-value and one sorted by gene name, in order to investigate not only the top significant SNPs but also any genes which had multiple significant intragenic SNPs.

The allele distribution between affected and unaffected individuals was examined for all significant SNPs for consistency with the known inheritance pattern. The genes within which SNPs were significant under the recessive model, and passed any of the other respective criteria, were included in the candidate geneset for further investigation.

2.3.1.3. Linkage analysis

For the linkage analysis, I chose to utilize the Merlin software package (Abecasis et al., 2002). Merlin is a widely used software package for conducting genetic linkage analyses on nuclear and extended pedigrees. Based on allelic inheritance, this approach was severely limited by the disconnected nature of the genotyped individuals in the flock. Because of this, the analysis was based on the only present nuclear family group (#398, #30, #RED1-4), one parent of which was not genotyped, which did not provide sufficient power to be used in candidate discovery. The linkage analysis was thus not able to be used in identification of candidate genes but instead to provide additional evidence for or against candidates identified by homozygosity mapping and association analysis.

Linkage analysis using the nuclear family (inclusive of #30, #398 and RED #1-4) with a parametric, recessive model, using a disease rarity of 0.0001, 0.0, 0.0, 1.0 (full penetrance, no penetrance effects) yielded a maximum possible LOD score of only 0.66, and a maximally significant p-value of 0.04.

The option ‘markerNames’ was used to report the specific markers involved rather than their map positions.

The options of ‘bits 40’ and ‘megabytes:25000’ were used to increase the available memory for the program in order to handle the complexity of the extended pedigree.

The ‘model’ option was included in order to run the analysis based on the parameters of the defined recessive model. The analysis was run with the following command:

```
$ merlin -d 48808SNPs.dat -p NuclearFamily.ped -m
3LineMapFile.map --markerNames --bits 40 --model
parametric.model --megabytes:25000 --swap >
LinkageResults
```

Attempted analysis of the wider genotyped pedigree was not informative.

2.3.1.4. Identification of the Candidate Geneset

Genes within the identified regions of homozygosity for affected individuals of the pedigree were included in the candidate geneset. Also included were genes within which a SNP was found to be significantly associated with disease using the association analysis. The limitations of the linkage analysis prevented the use of this approach to select SNPs for the candidate geneset, and they were instead used as supporting evidence to assess the selected candidates resulting from the association analysis and homozygosity mapping. A negative or neutral LOD score towards the SNPs within a particular gene candidate was counted against its credibility as a candidate, while a positive LOD score counted towards the same.

2.3.1.5. Biological Plausibility Investigation Candidates

All genes in the candidate geneset were then investigated using online resources and databases to find those which may be implicated in causing skeletal muscle-specific conditions.

The positioning of each SNP and gene was confirmed by accessing the online resource for the sheep genome v. 3.1.

Each gene under investigation was searched in the Online Mendelian Inheritance in Man (OMIM) database (McKusick, 2013) and in the Genecards database (www.genecards.org) for known gene expression profiles of the gene (e.g. was the gene

expressed in skeletal muscles), any biologically plausible mechanism by which it may have resulted in the disease, and for known disease associations which may fit the disease phenotype.

Google, Google Scholar and Pubmed were then interrogated for each candidate gene, and the following key words, in an attempt to find any relevant published data which might explain the disease phenotype: myotonic dystrophy; muscular dystrophy; nemaline myopathy; skeletal muscle; myopathy; dystrophy.

The best candidate from these complementary approaches was then chosen for follow-up using molecular biology.

2.3.2. Data Manipulation

2.3.2.1. SNP Array Data Manipulation

As the sheep reference genome was updated through the continued efforts of the ISGC the genomic positions of some SNPs was changed from previous iterations.

Large numbers of reported impossible recombination patterns between the offspring of #1563 and #1560 using pedstats and Merlin dictated the adjustment of the familial relationships then replace these individuals with ‘blanks’ (individuals with no reported values for SNP genotypes).

2.3.2.2. SNP validation

The genomic positions included in the provided .map files were based on the version 1.0 sheep reference genome. For analysis, these positions were updated to the version 3.1. genomic positions. This was carried out by extracting the 3.1. positions from the provided genome annotation file using a custom perl script and using PLINK to update the map and base pair chromosomal positions to the 3.1. reference version. This resulted in a number of SNPs changing position to unaligned scaffolds and the total number of SNPs in chromosomal positions on the reference genome to be 48 808.

2.3.2.3. PLINK

For association analysis the PLINK software package was utilised (Purcell et al., 2007, Purcell). For initial analysis, the adjusted files were modified to follow the PLINK format for missing values and uninformative individuals were trimmed from the analysis, but were otherwise unchanged.

2.3.2.4. Merlin

For linkage analysis the Merlin software package was utilised (Abecasis et al., 2002). The provided SNP data files were modified to follow the Merlin format using command line interface. The map files provided used chromosomal position and were 4 column files. Merlin requires the position in cM and 3 columns only. The cM positioning was approximated by dividing the chromosomal positioning by 1 000 000 and the extra column removed. Uninformative individuals were trimmed from the analysis.

2.3.2.5. Genome Sequencing

The provided genome sequencing files and annotation file had a different nomenclature for chromosomes than the reference FASTA sequence. The reference was modified using a perl script utilising the FASTA reader package.

The provided BAM files for individuals #398 and #1560 were realigned to the reference genome using BWA-MEM with the intention of improving the accuracy of alignment by using more recent tools than that used by the laboratory of Dr. James Kijas (Li, 2010). The alignment was validated using picard-tools and the validated SNP array (49 034) used as the database of known polymorphisms in order to put the realignment through the GATK workflow to call variants (McKenna et al., 2010, 2013). The lack of an indel database for sheep prevented taking these realignments through to a final file for analysis, and this path of investigation was abandoned in favour of using the alignments provided by Dr. James Kijas.

2.4. Molecular Biology

In order to strengthen the hypothesis of the prime candidate *ROCK2* as a causative disease gene for the OCPMD flock, additional investigation was undertaken at the molecular biological level. This included PCR amplification of the candidate gene cDNA and sequence analysis.

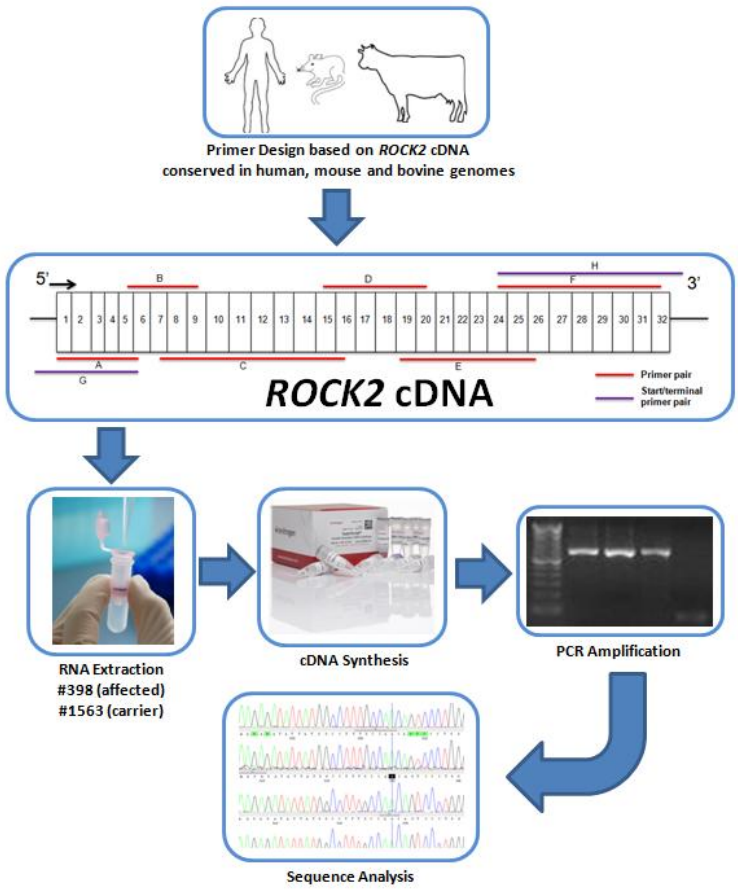


Figure 2.5. Flowchart for molecular biological investigation of prime candidate gene for the OCPMD pathology.

2.4.1. Primer Design

The genomic sequence at the position of *ROCK2* as reported by version 3.1. of the reference ovine genome was not adequately annotated and did not appear to match with the expected structure of *ROCK2*. Moreover the number of exons in the gene, being 32, and the size of the interspanning introns made screening this gene based on genomic DNA problematic. For this reason, the cDNA was targeted. The *ROCK2* cDNA sequence for sheep is not published.

Because of this, the conservation of the amino acid and nucleotide sequence across species was investigated using the software package Alamut v. 2.3, with NCBI build 36. While the amino acid sequence was found to be highly conserved across mammalian

species, the sequence of nucleotides was found to be only generally conserved in mammals.

Primers for this ovine study were therefore designed based on conserved regions identified within the *ROCK2* human, mouse and bovine cDNA sequences, as derived from the genome builds available in Ensembl (Flicek et al., 2013). The genome builds and links to each *ROCK2* cDNA sequence are provided in Table 2.1. Where 100% conservation between all three cDNA sequences was not possible, primers were instead designed in locations where 100% conservation existed between the bovine and human cDNAs, and minimal mismatches were identified in the mouse cDNA sequence. At the mismatch positions redundant nucleotides were used. These primer pairs (A-F, Table 2.2.) covered approximately 95% of *ROCK2*'s cDNA and were designed to be amplified at similar conditions to speed laboratory workflow. Additional primer sets (G & H, Table 2.2.) were designed at the terminus regions of the gene in order to attempt complete coverage of the cDNA sequence.

Table 2.1. Source of cDNA for ROCK2 human, mouse and cow species.

Species	**Genome Build**	**Web Link**
Human (*Homo sapiens*)	GRCh37	http://asia.ensembl.org/Homo_sapiens/Gene/Sequence?db=core;g=ENSG00000134318;r=2:11319887-11488456
Mouse (*Mus musculus*)	GRCm38	http://asia.ensembl.org/Mus_musculus/Gene/Sequence?db=core;g=ENSMUSG00000020580;r=12:16894978-16988274;t=ENSMUST00000020904
Cow (*Bos taurus*)	UMD3.1	http://asia.ensembl.org/Bos_taurus/Transcript/Exons?db=core;g=ENSBTAG00000005847;r=11:86501577-86583652;t=ENSBTAT00000007691

The primers were designed using the publically available Primer-BLAST tool (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) (Ye et al., 2012) and additionally validated using the Sigma Aldrich design tool (http://www.sigmaaldrich.com/configurator/servlet/DesignTool?prod_type=STANDARD) in order to ensure that the resulting oligonucleotides had minimal primer dimer and no more than a low chance of secondary structures forming.

Designed primer pairs were as follows:

Table 2.2. Primer sets designed for targeting of *ROCK2* cDNA.

Primer Pair	Equivalent exon location in cDNA	Primer Sequence* (5' – 3')	Estimated Product Size (bp)	Tm (°C)
A	1-6	F: CCCATCAACGTGGAGAG	637	60.3
		R: CAACTGCTGTATCYCAATGTACC		60.9
B	5-9	F: GCCTGGTGGAGACCTTG	630	62.0
		R: CTGCTGTCTATGTCACTGCTG		61.3
C	7-16	F: GTTTCCCTGAAGATGCAGAA	805	61.8
		R: CTTAACCGGGCTGCAGTA		61.4
D	15-20	F: AGTGAATCAACTCCAGAGACAA	819	60.4
		R: TGCTTTAACTGCTTTTCTGACA		61.5
E	19-26	F: ACAGAAAGATGTGCTAAATGAGG	792	61.1
		R: CGATTCATGATCTCAGCCA		62.5
F	24-32	F: CTAGTGATGTTGCCAATCTTG	1099	59.8
		R: ACCAACCGACTAACCCAC		59.9
G	Start codon - 6	F: **ATG**AGCCGGCCCCCG	777	71.2
		R: CAACTGCTGTATCYCAATGTACC		60.9
H	24 – 3'UTR	F: CTAGTGATGTTGCCAATCTTG	>1099	59.8
		R: GCWTTCAYAGAAGGCAG		60.1

*Redundant bases in red; Start codon of *ROCK2* in bold

2.4.2. RNA Extraction

Total RNA was extracted from the gracilus muscle (not affected in diseased individuals) of an affected (#398) and a carrier (#1563) sheep from the OCPMD flock. RNA was also extracted from the soleus muscle of the carrier but was not available from affected sheep as the soleus muscle is always affected (and hence has turned to fat) in aged diseased individuals. RNA extractions were performed using the RNEasy Fibrous Tissue Mini Kit (Qiagen) in accordance with the manufacturer's protocol (Qiagen, 2010). RNA concentration and quality (Abs. 260/280) was determined by a NanaDrop1000 (ThermoScientific).

2.4.3. cDNA synthesis

cDNA was synthesised from total RNA by reverse transcription using Superscript III First-Strand Synthesis System for RT-PCR (Life Technologies) and random hexamers (Invitrogen), following the manufacturer's protocol. The success of cDNA synthesis was confirmed by successful PCR amplifications of the *ROCK2* gene and electrophoresis.

2.4.4. PCR Amplification

For all amplification reactions the touchdown protocol was used (Korbie and Mattick, 2008). Each primer pair was initially trialed for PCR using the touchdown

program targeted to 63^{o}C, with one reaction containing and one not containing Q solution. Amplification of primer pair G was also attempted using a touchdown program targeted to 68 ^{o}C and a touchdown program targeted to 70 ^{o}C after failing to amplify at the touchdown 63 ^{o}C conditions. All PCR reactions were of 10 μL. Each reaction was as follows:

Table 2.3. PCR components for targeted amplicons within ROCK2

	With Q Solution	**Without Q Solution**
Component	**Volume (μL)**	**Volume (μL)**
H2O	4.6	6.6
10x buffer	1.0	1.0
5mM dNTPs	0.4	0.4
Forward Primer (50ng/μl)	0.4	0.4
Reverse Primer (50ng/μl)	0.4	0.4
Q solution	2.0	-
Taq polymerase (5 units/μl)	0.2	0.2
cDNA	1	1
Total	**10**	**10**

cDNA from the carrier's soleus and gracilus, and the gracilus of the affected sheep was used for PCR using all primer pairs. As the primers were designed to have 100% homology to the human sequence, for those primer pairs that failed under the initial PCR conditions, human cDNA (synthesised from RNA extracted from human skeletal muscle) was subsequently used as positive control to determine whether primer design or PCR conditions were the cause of failure.

2.4.5. Electrophoresis of PCR products

Unless otherwise specified, all PCR products were electrophoresed through a 1% agarose gel (containing 20 μL ethidium bromide / 400 mL agarose) with sodium borate buffer (0.4g NaOH /L; 2.4g H_2BO_3/1L; pH 7.0) using 3 μL of 4X loading buffer with 3 μL of reaction product, and size determined against a 100 bp ladder (Invitrogen).

2.4.6. Sequencing of Candidate Gene

PCR products were purified by pipetting through a purification tip (Diffinity RapidTip) for 1 minute. Sequencing reactions were then set up as follows:

Table 2.4. Sequencing reaction components for *ROCK2* PCR products.

Component	Volume (μL)
Big dye terminator mix	2.0
Sequencing buffer	2.0
Primer (forward OR reverse @ 50 ng/μl)	1.0
Template DNA	1.0-5.0*
Water	variable
Total	**10**

* Volume dictated by intensity of PCR products after electrophoresis. Strong intensity reduced the volume to 3 μL while a lesser intensity increased the volume to 5 μL

Table 2.5. Thermocycler protocol for sequencing reaction.

Step	Condition
1	96°C for 1 minute
2	96°C for 30 seconds
3	50°C for 30 seconds
4	60°C for 4 minutes
5	Goto step 2 24x

Each sequencing reaction was then taken through the following post-sequencing cleanup process:

1. Add 2.5 μL EDTA, 2.5 μL 3M sodium acetate, 25 μL 100% EtOH.
2. Vortex well and sit at room temperature for 15 minutes.
3. Centrifuge at 14000 rpm for 20 minutes.
4. Discard supernatant without disturbing pellet using pipette.
5. Wash well with 200 μL of 70% EtOH, dislodging pellet.
6. Centrifuge at 14000 rpm for 5 minutes.
7. Discard supernatant without disturbing pellet using pipette.
8. Dry in heating block @ 55 °C for at least 30 minutes or until dry.

The resulting samples were sent to the Lotterywest State Biomedical Facility: Genomics for sequencing.

2.4.7. Sequence Analysis

The sequencing results were aligned to the bovine cDNA sequence as reference for *ROCK2*, as derived from the Ensembl Cow release 73 (Flicek et al., 2013). Some sequencing results were not of sufficient quality to align to the reference and were automatically trimmed for quality at each end based on program default settings. This allowed most remaining sequences to be aligned to the bovine cDNA but several were still of insufficient quality to do so.

The aligned sequences were then analysed using the Codoncode aligner v.4.03 (CodonCode Corporation, 2013) in order to identify a possible homozygous mutation in the affected individual and a heterozygous mutation in the carrier. The base calls and chromatogram for the forward and reverse sequence of each PCR product was inspected for quality and divergence from the bovine cDNA reference or from consensus with the additional coverage of other strands. Where these diverged they were subjected to extra scrutiny.

3. Results

3.1. Summary

Bioinformatics methods were employed to find the best candidate gene/s for follow-up in the laboratory. For both homozygosity mapping and the association analysis, an initial analysis was carried out on 14 informative individuals and a subsequent final analysis carried out on 20 informative individuals (inclusive of the first sample group), when additional samples became available.

SNP homozygosity mapping was carried out to find homozygous regions of the genome in the affected individuals, while association analysis of the OCPMD SNP genotypes was conducted in order to determine statistically significant markers (and thus genes) of interest as potentially causative of the pathology.

Linkage analysis was performed using the SNP genotypes in order to provide additional evidence for any resulting candidates. While ideally we would utilize results from linkage analysis firstly, followed by association analysis only in those regions where linkage signals were detected thus maximizing statistical power and reducing the number of tests required, this was not possible due to lack of informative families (as noted earlier, and described further in Discussion). The candidate geneset was then investigated through interrogation of publically available databases and published literature for biological plausibility in order to identify best candidates for down-stream molecular biological analysis.

The *ROCK2* gene was identified as a priority candidate for further investigation using molecular biological techniques. Primers for cDNA amplification were designed in an overlapping manner based on conserved regions on the human, mouse and bovine cDNA sequences. RNA was extracted from gracilus skeletal muscle from an affected individual (#398) and from the gracilus and soleus skeletal muscle for a carrier sheep (#1563). cDNA was synthesized from all three RNA extractions and PCR amplification using various primer pairs was attempted for all samples. All successful amplifications were sequenced and the results analysed.

3.2. Results from Bioinformatics Analyses

Bioinformatics analyses included homozygosity mapping, association analysis and linkage analysis. These complementary approaches resulted in the selection of

approximately 50 to 100 candidate genes for further investigation from the published literature and publically available database resources.

3.2.1. Homozygosity Mapping

Homozygosity mapping was carried out using the SNP genotypes of the flock to determine regions of the genome that were homozygous in the affected individuals, as described in Section 2.3.1.1.

Initial Analysis

Table 3.1 indicates the longest regions of homozygosity observed in the initial dataset, derived from 14 informative individuals.

Table 3.1. Longest homoyzgous genomic regions observed in the SNP genotypes of the affected individuals in the initial dataset.

Location	N° sequential SNPs	N° affected*	N° carrier*	Length of run (bp)	Genes included
Start Chr3:19896460 *End* Chr3:20116222	8	9/9	0/5	219662	*PQLC3* *ROCK2*
Start Chr25:19167503 *End* Chr25:19438048	6	9/9	1/5	270545	*NRBF2* *JMJD1C*
Start Chr3:11661403 *End* Chr3:11916080	6	9/9	4/5	254677	*Uncharacterised*
Start Chr13:55866953 *End* Chr13:56010691	5	9/9		143738	*CDH26* *SYCP2*
Start Chr11:49904840 *End* Chr11:50045164	5	9/9		140324	*CCD57* *FASN* *LRR45*

* Sharing homozygous SNP genotype region

The run of homozygosity containing *ROCK2* was the longest observed in the initial dataset. The genes within the second and third longest runs of homozygosity were also put forward for additional investigation.

Final Analysis

Table 3.2. indicates the longest regions of homozygosity observed in the final dataset, derived from the 20 informative individuals available after additional SNP genotyping.

Table 3.2. Longest homozygous genomic regions observed in the SNP genotypes of the affected individuals in the final dataset.

Location	N° sequential SNPs	N° affected*	N° carrier*	Length of run (bp)	Genes included
Start Chr3:19989596	6	14/14	2/6	126626	PQLC3
End Chr3:20116222					ROCK2
Start Chr25:19270565	5	14/14	2/6	167483	JMJD1C
End Chr25:19438048					
Start Chr11:49904840	5	14/14	5/6	140324	CCDC57
End Chr11:50045164					FASN
					LRRC45
Start Chr13:55866953	5	14/14	6/6	143738	CDH26
End Chr13:56010691					SYCP2

* Sharing homozygous SNP genotype region

The longest observed run of homozygous SNPs for affected individuals in the final dataset covered 6 sequential SNPs, encompassing 126626 bp. This region, covering SNP 3 – 6 (Table 3.3) while shorter than that seen in the initial dataset, covering SNP 1 – 8 (Table 3.3), reassuringly still included *ROCK2*, suggesting its viability as a potential candidate of interest. In this larger dataset, with the shorter region of homozygosity, there were 2 carriers out of 6 who shared the homozygous SNP genotypes with the 14 affected. The next longest run of homozygous SNPs observed were of 5 markers in length. There were 2 of these regions observed and all genes involved put forward for further investigation.

The SNP genotypes for all informative individuals of the flock for the *ROCK2*-containing homozygous SNP region were extracted from the SNP genotype pedigree file using the custom perl script "Haplotype_Extractor" (see appendix VI for source code).

Table 3.3. SNP genotypes of the affected individuals within the ROCK2-containing run of homozygosity

ID	Father	Mother	Sex*	1	2	3	4	5	6	7	8
398	1565	1564	2	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
1242	1137	1112	1	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
1406	1137	1108	1	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
1409	1137	1111	2	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
1279	1103	1128	1	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
1503	1103	1107	1	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
1508	1103	1105	2	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
1512	1103	1134	2	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
1518	1103	1132	2	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
Red1	30	398	1	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
Red2	30	398	1	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
Red3	30	398	1	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
Red4	30	398	2	G/G	G/G	G/G	A/A	G/G	G/G	A/A	G/G
74	1560	1564	1	G/G	A/G	G/G	A/A	G/G	G/G	A/A	G/G

*1 Male; 2 Female

Table 3.4. SNP genotypes of the carrier individuals within the *ROCK2*-containing run of homozygosity

ID	Father	Mother	Sex*	SNPs+	1	2	3	4	5	6	7	8
1560	130	1512	1		A/G	A/G	G/G	G/A	A/G	A/G	A/A	G/G
1561	131	1508	2		G/G	A/G	G/G	A/A	G/G	G/G	A/A	G/G
1563	131	2001	2		G/G	A/G	G/G	A/A	G/G	G/G	A/A	G/G
1107	0	0	2		G/G	A/G	G/G	G/A	G/G	A/G	A/G	G/G
1119	0	0	2		G/G	G/G	A/G	0/0	G/G	A/G	A/G	G/G
207	1560	1563	2		A/G	A/G	G/G	G/A	A/G	A/G	A/A	G/G

*1 Male; 2 Female; +highlighted SNPs represent those homozygous in carriers #1561 and #1563

The SNP genotype observed for SNP 2 in #74 (highlighted in yellow) was responsible for shortening the length of the homozygous region in affected individuals of the flock. The inheritance of this allele from #1560 and #1563 is consistent with this result.

3.2.2. Association Analysis

Genome-wide association analysis was conducted on the SNP genotypes of the flock, using PLINK, to determine whether SNPs which showed statistically significant correlation ($P<0.05$) to the disease state under a recessive inheritance model, as described in Section 2.3.1.2.

Initial Analysis

The initial analysis found 371 SNPs which met this significance level and were within a gene. Of these, 5 were x-linked and thus excluded due to the autosomal inheritance pattern. Being a slightly larger sample size and thus more informative, the final analysis, inclusive of the individuals of the initial dataset, found 415 SNPs which met the 0.05 significance level. Of these 4 were x-linked and were thus excluded. The top 20 SNPs by p-value are included Table 3.5., as is the SNP within *ROCK2* also found to be significant (reported at the bottom of the table).

Table 3.5. Top SNPs by significance for initial association analysis.

Rank	SNP	Gene	'affected' genotype for affected*	'affected' genotype for unaffected*	p-value
1	OAR12_83564723.1	Uncharacterised	8/1	0/5	0.001281
2	OAR2_195165011.1	PTPN4	9/0	1/4	0.001499
3	s42749.1	MAN2A1	7/0	0/3	0.001565
4	s53402.1	FTCD	0/9	3/1	0.003054
5	s09120.1	PPP1CC	6/1	0/5	0.003415
6	OAR1_135764540.1	TAK1L	7/2	0/5	0.005289
7	OAR1_38581202.1	DOCK7	7/2	0/5	0.005289
8	OAR2_219701580.1	ADAM23	7/2	0/5	0.005289
9	OAR1_65172267.1	SYDE2	0/9	3/2	0.008752
10	OAR10_49941391.1	KLF12	9/0	2/3	0.008752
11	OAR12_82572003.1	C1ORF53	0/9	3/2	0.008752
12	OAR14_29971688.1	CDH8	0/9	3/2	0.008752
13	OAR17_45018430.1	GRIA2	0/9	3/2	0.008752
14	OAR23_36959489.1	GREB1L	0/9	3/2	0.008752
15	OAR23_59623047.1	Uncharacterised	0/9	3/2	0.008752
16	OAR3_123897974.1	PTPRQ	0/9	3/2	0.008752
17	OAR8_56910466.1	THEMIS	0/9	3/2	0.008752
18	OAR8_58134022.1	LAMA2	0/9	3/2	0.008752
19	OAR8_58181648_X.1	LAMA2	0/9	3/2	0.008752
20	s03652.1	GLRB	0/9	3/2	0.008752
	…	…	…	…	…
=200	OAR3_21684794.1	ROCK2	9/0	3/2	0.04042

*'Affected' genotype wherein tested alleles are homozygous

Table 3.6. Association analysis results of SNPs within the ROCK2 gene in the initial dataset.

Chr	SNP	Alleles	Disease Model	Numbers affected	Numbers carrier	Chi2	df	p-value
3	OAR3_21684794.1	A G	Recessive	9/0	3/2	4.2	1	0.04042
3	OAR3_21630699.1	A G	Allelic	0/18	3/7	6.048	1	0.01392

OAR3_21684794.1 was found to be significant within the initial dataset at a p-value of 0.04042, ranking at =200 of 366 significant SNPs, with all 9 affected individuals possessing the AA alleles (Table 3.6.).

Final Analysis

After additional genotyping, the new dataset provided further results, changing the significance of a number of SNPs and causing a number to become less significant than they were in the initial dataset. Table 3.7 summarises the top 20 SNPs by p-value, along with the SNP within *ROCK2*.

Table 3.7. Top SNPs by significance for final association analysis.

Rank	SNP	Gene	'affected' genotype for affected*	'affected' genotype for unaffected*	p-value
1	OAR12_83564723.1	Uncharacterised	12/2	0/6	0.000336
2	OAR2_195165011.1	PTPN4	14/0	2/4	0.000636
3	OARX_29188769.1	Uncharacterised	11/3	0/6	0.001209
4	s43015.1	CCL26	11/3	0/6	0.001209
5	OAR3_193744872.1	IFT27	10/3	0/6	0.001799
6	s10237.1	IL21R	10/3	0/6	0.001799
7	s09120.1	PPP1CC	10/1	1/5	0.002205
8	OAR10_33307197.1	Uncharacterised	12/2	1/5	0.00301
9	OAR10_69032148.1	GPC5	12/2	1/5	0.00301
10	OAR10_25988324.1	NBEA	10/4	0/6	0.003415
11	OAR24_30139569_X.1	MRPS17	10/4	0/6	0.003415
12	OAR24_30146533.1	MRPS17	10/4	0/6	0.003415
13	OAR3_19616550_X.1	KIDINS220	10/4	0/6	0.003415
14	OAR16_71436671.1	ADCY2	0/14	3/3	0.004108
15	OAR17_45018430.1	GRIA2	0/14	3/3	0.004108
16	OAR17_75701731.1	INPP5J	0/14	3/3	0.004108
17	OAR18_16433946.1	AGBL1	0/14	3/3	0.004108
18	OAR23_36959489.1	GREB1L	0/14	3/3	0.004108
19	OAR3_35192406.1	DTNB	0/14	3/3	0.004108
20	OAR4_26645098.1	TSPAN13	0/14	3/3	0.004108
	…	…	…	…	…
=128	OAR3_21684794.1	ROCK2	14/0	4/2	0.02278

*'Affected' genotype in which both tested alleles were homozygous

Of the 411 included significant SNPs in the final dataset, 158 SNPs also appeared in the results of the initial dataset and were found to improve in significance by approximately 50 %. Twenty nine of the SNPs in the final analysis were found to be less significant than in the initial dataset and were thus excluded. Six were only slightly more significant (compared to the 158 which were in general ~twice as significant) and excluded. This provided 376 SNPs of significance which occurred within a gene and were found to be significantly associated with the disease, 158 of these being considered part of the 'overlap set'. All of the genes in which these SNPs occurred were placed into the candidate geneset for further investigation, and the overlap set given additional scrutiny.

The statistically significant SNP within *ROCK2* was found to be of strengthened significance in the larger dataset, showing improved numerical p-value (by approximately 50 %) and relative significance to the other SNPs tested. In the initial dataset OAR3_21684794.1 was ranked equal 200th of 366 SNPs (<0.05; placing it in the top ~50 % of significant SNPs). In the final dataset OAR3_21684794.1 was ranked

equal 128th of 411 SNPs (<0.05; placing in the top ~30 % of significant SNPs). Given the limitations of the analyses, the exploratory nature of the investigation and a p-value of 0.02278 (Table 3.9) this was taken as evidence for the potential involvement of this region with OCPMD.

Table 3.8. Association analysis results of SNPs within the ROCK2 gene in the final dataset.

Chr	SNP	Alleles	Disease Model	Numbers affected	Numbers carrier	Chi2	df	p-value
3	OAR3_21684794.1	A G	Recessive	14/0	4/2	5.185	1	0.02278
3	OAR3_21630699.1	A G	Allelic	0/28	4/8	10.37	1	0.001281

3.2.3. Linkage Analysis

Linkage analysis was conducted on the SNP genotypes of the available flock members to determine regions of the genome which may show evidence of linkage to the disease state, using the program Merlin as described in Section 2.3.1.3.

Linkage analysis yielded a maximum LOD score of only 0.66, and a minimum p-value for this of 0.04. There were 1614 SNPs which had a LOD of 0.66 and p-value of 0.04. Because of this, the linkage analysis was deemed to be too uninformative (there weren't enough complete nuclear family groups) to alone convincingly indicate possible target genes for further investigation, and was thus only used as additional evidence for or against any potential targets.

The analysis of SNPs within a region of chromosome 3, which included *ROCK2*, yielded the following results:

Table 3.9. Linkage analysis for SNPs within the *ROCK2*-containing homozygous region

SNP	LOD	p-value
s04366.1	0.54	0.06
s43586.1	0.53	0.06
s44598.1	0.51	0.06
s62248.1	0.51	0.06
s39730.1	0.50	0.06
OAR3_21684794.1	0.48	0.07
OAR3_21630699.1	0.50	0.07
OAR3_21695741.1	0.47	0.07

While none of these SNPs reached the maximum LOD score possible for the dataset, the positive LOD scores in the region surrounding *ROCK2* provided supporting evidence for potential involvement of the gene with the disease, and thus further

supports it as our prime candidate, especially in view of negative LOD scores for many other candidates in consideration.

3.2.4. Whole-genome sequencing

The whole-genome sequencing data provided by Dr. James Kijas was originally intended to be used for discovery, but on inspection found to be of too low coverage (8x) to provide sufficient information for this purpose, having been sequenced for assembly purposes. For use in a project like this one the coverage would ideally be ~100x and so this line of investigation was abandoned in favour of the other approaches.

3.2.5. Candidate Geneset

The candidate geneset for biological investigation comprised the candidates selected from homozygosity mapping and the association analysis, with any genes of potential interest also examined in regards to the results of linkage analysis.

Of the three SNPs occurring within the *ROCK2* gene, one was able to be tested under the recessive model. One was not able to be tested under the recessive model due to a lack of divergent SNP genotypes, but was able to be tested under an allelic model, and results were suggestive of a significant effect (Table 3.8). The remaining SNP was of the same genotype for all individuals in the sample group and thus could not be tested.

3.2.6. Biological plausibility investigation

Between 50 and 100 genes were chosen for further investigation based on evidence from homozygosity mapping, association analysis, and supported by linkage analysis. The best candidate was concluded to be *ROCK2*.

ROCK2 is a rho-associated kinase, the predominant ROCK isoform in skeletal muscle. Rho-dependent kinases like *ROCK2* are major downstream effectors of RhoA, important in regulating myogenesis, however the specific functions of the ROCK isoforms are not currently well defined. It is progressively upregulated during myoblast differentiation, and an isoform of *ROCK2*, *ROCK2m*, has been shown to be preferentially expressed in skeletal muscle. This isoform is generated by alternative splicing of an evolutionarily conserved intron occurring between exons 27 and 28, called 27’ (Pelosi et al., 2007). Rock2m was demonstrated to be common to human, rat and mouse genomes and is heavily upregulated during myogenic differentiation (Pelosi

et al., 2007). This information suggests that a mutation within this intron, downregulating expression of the Rock2m isoform or changing the protein product, could result in a skeletal muscle-specific pathology.

3.3. Molecular Biological Investigations of Prime Candidate Gene

The best candidate chosen from bioinformatics analyses, *ROCK2*, was investigated in the laboratory using molecular biological techniques. The goal of this work was to sequence the cDNA from an affected and carrier sheep, in pursuit of identifying a putative disease-causing variant that was homozygous in the affected and heterozygous in the carrier. The region around and including the *ROCK2m* isoform-specific intron, 27', was considered to be the most likely location of such a mutation.

3.3.1. RNA Extraction

RNA was extracted from skeletal muscle tissue of an affected (#398) and a carrier (#1563) sheep of the OCPMD flock in order to synthesise cDNA for subsequent PCR amplification and sequencing for analysis. The muscles with the most pathology in the affected individuals (e.g. soleus) were unlikely to provide sufficient muscle tissue for analysis (as they had turned to fat) so a muscle not normally affected (gracilus) was chosen for RNA extraction. In the carrier, an RNA extraction was carried out for both the gracilus and the soleus. A second elution was performed though the RNA columns to ensure all RNA was recovered.

Table 3.10. RNA extraction concentrations from skeletal muscle tissue.

Tissue	**Elute**	**RNA conc. (ng/μL)**	**Absorbance 260/280***
398 Gracilus	1	29.55	1.84
	2	20.00	1.815
1563 Gracilus	1	219.8	1.98
	2	52.95	1.96
1563 Soleus	1	82.4	2.045
	2	20.1	1.98

RNA was successfully extracted from all tissues. The extraction from the affected skeletal tissue was of considerably lower concentration than the other samples, and the absorbance ratio at 260/280 nm indicated a less pure sample, however it was still considered to be of reasonable quality after checking it by gel electrophoresis and primer amplification. The extractions from the carrier gracilus and soleus were of higher quality and quantity.

3.3.2. cDNA Synthesis, PCR Amplification & Electrophoresis

cDNA was synthesized from the affected and carrier RNA samples. PCR amplification of each primer pair was attempted using each of the cDNA samples and the product purity and size confirmed by electrophoresis.

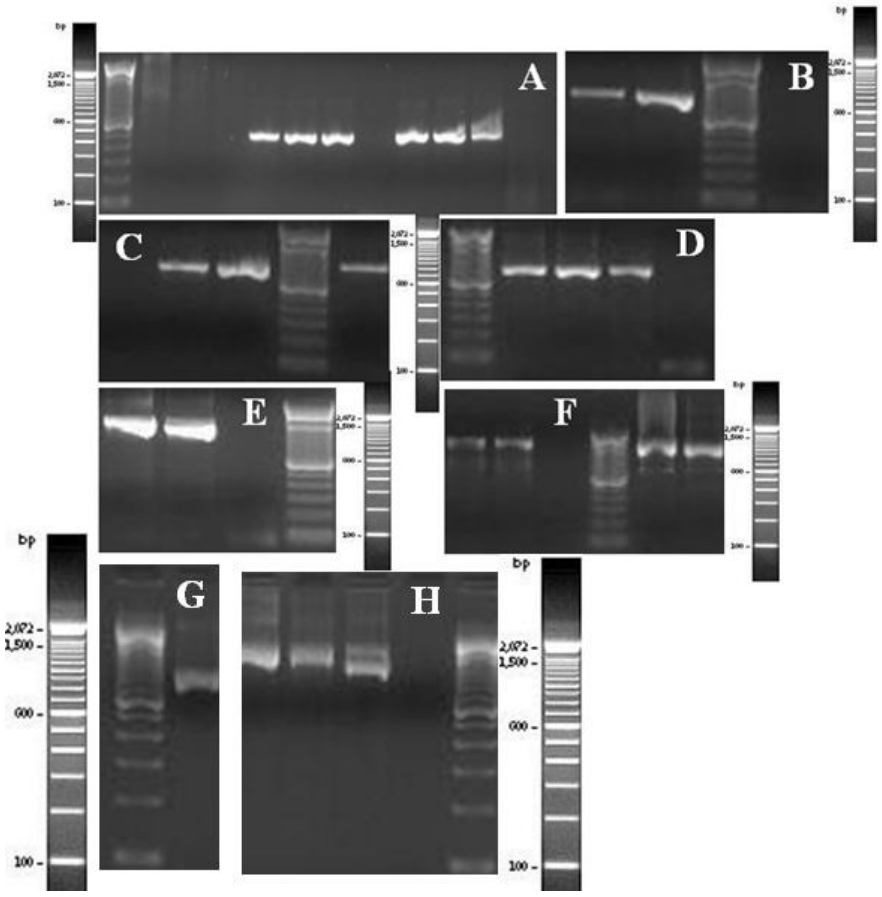


Figure 3.1. Electrophoresis of PCR products from *ROCK2* primer pairs. Gel columns are numbered sequentially from 1, left to right. Affected gracilus is referred to as (1) carrier gracilus as (2) and carrier soleus as (3), while the amplifications of control human skeletal muscle are (4) and (5). PCR products as follows: **Gel A.** 1: 100 bp ladder; 5: 1B; 6: 2B; 7: 3B; 8: H_2O; 9: 1A; 10: 2A; 11: 3A; 12: H_2O. **Gel B.** 1: 1C; 2: 2C; 3: 100 bp ladder; 4: H_2O. **Gel C.** 1: H_2O 2: 1D 3: 2D 4: 100 bp ladder 5: 3D **Gel D.** 1: 100 bp ladder 2: 1E 3: 2E 4: 3E 5: H_2O **Gel E.** 1: 3F 2: 5F 3: H_2O 4: 100 bp ladder **Gel F.** 1: 2H 2: 3H 3: H_2O 4: 100 bp ladder 5: 4H 6: 5H. **Gel G**. 1: 100 bp ladder 2: 3C **Gel H.** 1: 1F 2: 1FQ 3: 5F 4: H_2O 5: 100 bp ladder

Table 3.11. Primer sets amplified to the predicted size based on electrophoresis of PCR products

Primer Set	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Affected gracilus	✓	✓	✓	✓	✓	✓	X	
Carrier gracilus	✓	✓	✓	✓	✓		X	✓
Carrier soleus	✓	✓	✓	✓	✓	✓	X	✓

Gel electrophoresis showed expected product sizes (Table 3.11.) for most primer sets amplified through PCR. While the carrier gracilus cDNA did not have a clear gel

image for its PCR product for fragment F its subsequent sequencing demonstrated its successful amplification.

3.3.3. Sequencing

Sequencing of most *ROCK2* primer sets was successfully carried out and aligned to the bovine cDNA as reference, due to the unavailability of the ovine *ROCK2* cDNA sequence.

Table 3.12. PCR products that were successfully sequenced and aligned to the bovine *ROCK2* cDNA reference.

Primer Set	**A**		**B**		**C**		**D**		**E**		**F**		**G**		**H**	
Direction	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**
Affected gracilus	✓	✓	✓	✓	✓	✓				✓						
Carrier gracilus	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Carrier soleus	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				

Primer sets which covered the termini of *ROCK2* were not able to be successfully amplified or the resulting PCR products sequenced and, while the electrophoresis of D, E and F primer sets for the affected indicated a PCR product of the expected sizes, these were not able to be successfully sequenced and aligned to the bovine cDNA.

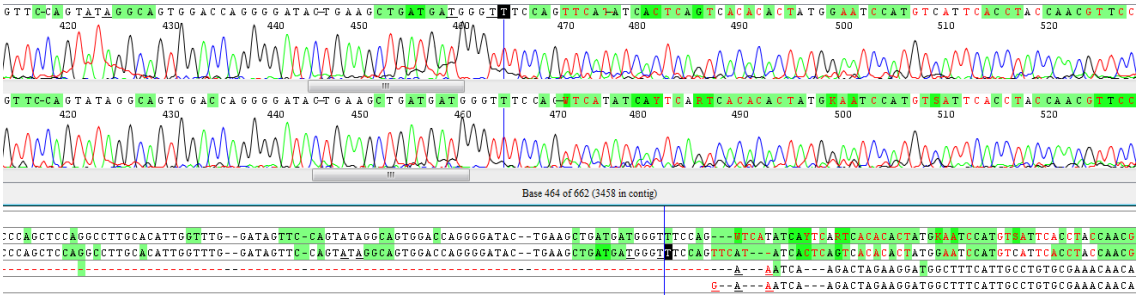
The total size of the bovine *ROCK2* cDNA is 4101 bp and this sequence was chosen to be the reference sequence upon which to attempt to align the sequencing results from the PCR products derived from the ovine cDNA samples. After quality control based on examination of the individual chromatograms for each sequencing result, there were 36 differences between the consensus ovine sequence results and 3920 bp of the bovine cDNA sequence covered, some of these resulting in amino acid changes.

Of note, the forward strand sequences for primer set F for both the carrier gracilus and soleus appeared to mismatch with the bovine cDNA sequence, but have 100% agreement with one another. The sequences of the reverse strand were in near perfect agreement with the bovine cDNA reference, however. It was determined that these PCR products included a transcript that contained an additional 171 bp to the bovine cDNA reference sequence. When comparing these nucleotides to that of the bovine genomic sequencing for *ROCK2*, the 171 bp (equating to 57 amino acids), represent the inclusion of a section of intron 27 of the gene, relating to the muscle-specific exon in the *ROCK2m* transcript. The bovine and ovine sequences for these 171 bp is identical, with

there being 8 nucleotides different between the bovine/ovine and human cDNA sequences for this region.

Of the regions that were covered by sequencing of good quality, there were no obvious variants that were homozygous in the affected and heterozygous in the carrier, indicative of a disease-causing variant, e.g. a mutation.

Figure 3.2. Snapshot of the sequencing chromatogram of a segment of fragment F highlighting the beginning of the variable region of the ROCK2 transcript.



There is a drop in relative intensity of the chromatogram peaks at the approximate start of the putative 27' (indicated by the vertical line). There is also an increase in background/noise from this position onwards, due to the presence of two divergent transcripts.

4. Discussion and Future Directions

This work has identified a prime candidate gene, *ROCK2*, which may contain the causative mutation of OCPMD. Further, it has additionally successfully sequenced much of the cDNA for *ROCK2* and identified a novel alternative transcript comparable to *ROCK2m*, previously only reported in humans, mice and rats. This is the first evidence for the existence of *ROCK2m* in sheep, the transcript resulting from the inclusion of an additional exon (27') derived from a conserved intronic region between exons 27 and 28 being spliced into the transcript. Additionally, this work has set the stage for the complete sequencing of this transcript in the OCPMD flock in search of the putative disease-causing mutation.

Despite substantial past efforts over the past two decades towards revealing the genetic cause of OCPMD, traditional methods of investigation were unsuccessful. The major hurdles obstructing these investigations, such as an incomplete and thus poorly-annotated sheep genome, have been mostly overcome. The advent of SNP arrays, second-generation sequencing and the work of the ISGC, along with advancements in bioinformatics tools and techniques, have provided the best opportunity yet to discover the causative disease gene, despite the continued existence of some limitations. The collaboration with Dr. James Kijas at CSIRO has provided a conduit to the ISGC and a means by which SNP genotype and whole-genome sequencing data could be obtained. Another boon to this research was the birth, in 2012, of 4 sibling lambs affected by OCPMD, through *in vitro* fertilisation techniques. The inclusion of their genetic material into the analyses added greatly to their informativeness.

Bioinformatics analyses, comprising three separate lines of investigation, pointed towards *ROCK2* as a plausible candidate for a causative mutation in the OCPMD flock. While some other candidates were of greater statistical significance in one particular approach, none besides *ROCK2* was implicated by all three.

Using the homozygosity mapping as the starting point for our analyses overcame many of the problems associated with the small dataset, as this data filtering process was not subject to statistical assumptions. Of all genes within regions implicated by homozygosity mapping, only the region containing *ROCK2* contained SNPs which met the nominal 0.05 significance level in association analysis. Additionally, these SNPs were found in linkage analysis to have positive LOD scores associated with them, further implicating ROCK2. As with the *ROCK2*-containing homozygous region, the

next most interesting candidate gene from homozygosity mapping, *JMJD1C*, shared its 'background homozygosity' (that of the SNPs) with 2 of 6 carriers. For these homozygous SNPs, however, linkage analysis provided negative LOD scores, and none of these SNPs achieved significance in the association analysis.

In the case of the *ROCK2*-containing region, the SNP genotyping for one of carriers homozygous for the run of SNP homozygosity (#1563) has some possible issues of data quality, as there were a large number of Mendelian inheritance errors reported due to 'impossible recombination patterns' registered during linkage analysis**.** In addition, both carriers who shared this background homozygosity were bred from the same sire, a normal individual from outside of the pedigree, which may have resulted in a background SNP haplotype the same as that of the affected individuals without the putative causative mutation (it may have arisen only relatively recently on the same background haplotype).

In the case of the *JMJD1C*-containing region, there is no explanation yet for this sharing of 'background homozygosity' with the affected individuals. As mentioned previously however, the sharing of background homozygosity with carriers of the flock was not alone considered sufficient evidence to disqualify the genes within a region of homozygosity in the affected individuals from investigation.

The association analysis resulted in over 300 SNPs which met the criteria for inclusion in the candidate gene set, and the genes within which these were located (or were physically close to) were further investigated. The 158 SNPs of the 'overlap set', in becoming more significant in the final analysis, were (ranked) higher up our list of potentials, as were the genes within which were located any of the 100 most highly significant SNPs by p-value, and any genes in which multiple SNPs achieved statistical significance. The allele distribution was also taken into consideration (as the 'control' sheep in the analysis were all known carriers), as was the fact that a significant SNP might not itself be the causative mutation, but instead simply be in strong LD with the variant. It was considered that in this discovery-based research project it was greater importance to cast a wider net, as there is no replication dataset or 'gold standard' for comparison analysis (e.g. no dbSNP for sheep). The situation is akin to that of the 1990's for researchers studying human diseases. In the case of replication of previous work or two independent datasets, more stringent cutoffs would have been employed.

While many hurdles and challenges were overcome to make this work possible, additional remained that limited or restricted some of the bioinformatics analyses. Without the valuable contributions of Dr. James Kijas in performing the whole-genome sequencing and SNP genotyping it would not have been possible to conduct the homozygosity mapping, association analysis or linkage analysis for this project. However, due to the limited reagents for the SNP array still available (the SNP array has been discontinued) only a subset of the sheep for which we had biological samples available were able to be SNP genotyped and thus the pedigrees were not as informative (complete) as would ideally be possible. In addition, the lack of a SNP database beyond the SNP array, HapMap sheep equivalent (or Merino-specific database), dbSNP for sheep, or well defined reference genome made accounting for any larger population-based background genetics difficult.

The association analysis was inhibited by a smaller sample size than would have been ideal, and a true GWAS would have required SNP genotyping of much finer detail genome coverage, though this will be possible in the future as the sheep genome project matures. The fact that the tests did not meet the criteria for independence and small number of individuals prevented traditional control for multiple testing, as many of the multiple-testing correction methods are for independent samples (Gao et al., 2010), however this was overcome by using our complementary, overlapping analysis approaches.

The linkage analysis was limited by not only the complexity of the pedigree (creating computational hurdles with large memory requirements) but also by the limitations on which individuals were able to be SNP genotyped. This resulted in a pedigree for which the large majority of individuals genotyped were not connected into nuclear families (integral to map inheritance through linkage analysis), thus reducing the informativeness and power of the dataset. This led to significant time being expended fruitlessly and a number of strategies being attempted and discarded. These included the use of alternate tools such as LAMP (Li et al., 2005) and strategies to maximise potential results, including carrying out chromosome-specific linkage analysis, prior to the additional genotyping of connected individuals (#398, #30, RED #1 - RED #4) allowing the use of Merlin to conduct a limited linkage analysis. This analysis was most likely underpowered, as evidenced by the distribution of LOD scores across the genome (Appendix IV) but was still helpful in building candidate genes for further investigation.

Despite some of the limitations, it was reassuring that the prime candidate chosen for follow up, *ROCK2*, was supported by all lines of investigation; homozygosity mapping, association analysis, and linkage analysis. It was additionally considered through interrogation of the literature to represent a biologically relevant and plausible candidate for involvement in the OCPMD phenotype.

Implications

Skeletal muscles are essential for the functions of daily life; required for breathing, movement, swallowing and walking. As outlined in the introduction, skeletal muscle fibres can be classified into slow twitch or fast twitch, with different muscles in the body comprised of varying proportions, based on their functions.

As discussed in Section 1.4., diseases of skeletal and cardiac muscles can be emotionally, socially and economically devastating. Some muscle diseases are lethal before or shortly after birth. Progressive diseases like OCPMD have a daily impact and disable patients over a lifetime. Unfortunately for patients and their families, very little in the way of treatments are available, although supportive measures such as ventilators and wheelchairs can mitigate the impact of symptoms.

While the first step to rationally devising therapeutic approaches for a disease is information of the genetic causes, identification of the disease gene is not always sufficient in itself to design possible treatments. It is crucial to understand how the mutated gene leads to disease at the DNA, RNA and/or protein levels and through to the pathobiology of the disease. This can be a complex problem to solve, and one of the best ways to resolve this complexity is through animal models (Abresch et al., 1998). Moreover, an animal model is often essentially required as a test-bed for evaluating potential therapies.

Rather than generating a new mouse, zebrafish or *Drosophila* animal model (the more typical laboratory animal for genetic studies) of a particular human genetic muscle disease, this project has played a vital role in revealing the likely genetic cause of an existing, internationally unique, sheep model of a muscle disease. This is similar in approach to projects that phenotype mice subjected to ENU mutagenesis and subsequently attempt to identify the causative mutation in phenotypically interesting results (Fossett et al., 1990). The advantage of this project is that we have already identified the OCPMD flock as of major phenotypic importance in regards to human

disease and know that the mutation is naturally occurring in the pedigree, suggesting that it could well have an analogue in humans.

For decades, this Western Australian sheep model has been considered of high value to the medical research community, in that the exhibited pathologies in skeletal muscles of affected sheep overlap with different human skeletal muscle diseases, as discussed in Section 1.6.1. Additionally, the selective nature of slow fibre-predominant muscles being affected is reminiscent of some human skeletal muscle diseases such as McArdle's disease (Miteff et al., 2011). Lastly, the size of sheep, along with their skeletal muscle fibre-predominance and distribution is comparable to humans, especially in contrast with most existing animal models of muscle disease, such as mice. One of the biggest hurdles in potential treatment for disorders affecting the skeletal muscles is delivery to the target cells (Fairclough et al., 2013). Thus, if a potential therapy were to be trialled in the OCPMD sheep and found to be successful, it might be presumed that this therapy would have a greater chance of being effectively translated to humans as opposed to a treatment proven in mice.

ROCK2 is the predominant predominant *ROCK* isoform expressed in the skeletal muscles which, from first principles, must be the case in any gene causing a skeletal muscle-specific pathology (Pelosi et al., 2007). *ROCK2* is a downstream effector of Rho, which has been implicated in the control of skeletal muscle differentiation (Doherty et al., 2011). Not supporting this hypothesis, however, was the fact that the major isoform of *ROCK2* is also expressed in a range of other tissues, such as the heart, lung and brain, none of which are affected in OCPMD (Pelosi et al., 2007, discussed in Section 1.5.3). The discovery of a *ROCK2* isoform, *ROCK2m*, which is preferentially expressed in the skeletal muscles and heavily upregulated during myogenic differentiation, provided a possible mechanism by which a mutation in *ROCK2* could result in a pathology limited to skeletal muscles (Pelosi, et al., 2007). However, for completeness, the whole cDNA sequence was attempted to be sequenced in the skeletal muscles from both the carrier and affected sheep as it was also possible that a variant elsewhere in the gene could be disease-causing and produce exclusively a muscle phenotype. For example, a missense variant might exist in exon 3 of *ROCK2*, and produce an amino acid change in all transcripts. If this amino acid is involved in binding another protein that is only present in skeletal muscle, then such a conversion may only have consequence in skeletal muscle.

As described earlier, *ROCK2m* is an alternative isoform of *ROCK2* which is expressed preferentially in the skeletal muscle tissues. Previously demonstrated only in humans, mice and rats, *ROCK2m* is differentiated from the major *ROCK2* isoform by inclusion of a highly evolutionarily conserved intronic region. This is located between exons 27 and 28, and dubbed 27' (Pelosi et al., 2007). In mice, 27' is 171 bp and its inclusion in *ROCK2m* results in a transcript 57 amino acids longer than the major isoform. Thus it became of biological plausibility that a variant within this 171 bp region, a deletion within or encompassing it, or a change affecting the splicing of 27' or *ROCK2m*, could result in a skeletal muscle-specific phenotype.

In designing a method for screening *ROCK2* for a potentially causative variant, it would have been ideal to target 27' directly. Unfortunately, due to the incompleteness of the ovine reference genome, and the low level of coverage (8x on average) afforded by the whole-genome sequencing of #398 and #1560, it was not possible to design primers to target genomic DNA in the surrounding regions of 27'. Despite the highly evolutionarily conserved nature of the 27' sequence in mice, humans and rats, the flanking nucleotides were not, making primer design for ovine genomic DNA from the sequences of other species difficult and unlikely to be successful. In addition, the amplification and sequencing of 27' alone would have ignored the possibility of a causative variant within the major *ROCK2* isoform; the muscle specificity related to a specific binding partner present in affected tissues.

ROCK2 in sheep is 140,764 bp long. The coding regions of *ROCK2* were found to be highly conserved across mammalian species, but the introns were not. This, combined with the limitations of the sheep reference genome, suggested that to design primers based on the genomic DNA flanking exons would be unlikely to find success. The lack of a published sheep *ROCK2* cDNA sequence and conservation of coding regions across species, in combination with the relatively large number of exons (32) spanning approximately 4 kbp, suggested that the most viable strategy would be primer design based on highly conserved regions in human, mouse and bovine cDNA sequences in order to sequence the entirety of the *ROCK2* cDNA. Therefore, we decided that this was best accomplished by the amplification and sequencing of the cDNA in overlapping fragments rather than amplifying and sequencing each exon from genomic DNA individually.

The presence of 27' and *ROCK2m* was demonstrated in humans and mice previously, and the bovine genome is the closest to that of sheep published. Thus these three cDNA sequences were chosen for primer design. These were aligned and primers designed in overlapping pairs based on regions of 100% of near-100% identity. Eight primer pairs were created so as to produce overlapping fragments encompassing the entirety of the *ROCK2* cDNA sequence. While it would have been preferable to have designed a forward primer in the 5' UTR of the cDNA sequence rather than the start codon, the 5' UTR sequence was not available for the bovine *ROCK2* cDNA sequence, and that of the human and mouse *ROCK2* cDNA showed variability between one another, indicating that a primer designed for the ovine 5' UTR would be unlikely to find success. Because of this limitation, it is probable that a disease-causing variant in the first 15 nucleotides of *ROCK2* would not be identified by my analyses as, for these bases, the sequence of the primer would be observed in the sequencing results, rather than that of the cDNA.

Each cDNA sample created from the extracted skeletal muscle RNA was successfully amplified by multiple primer pairs, indicating that the extracted RNA was of sufficient quality, and the cDNA synthesis efficacious. Fragment G was the only fragment not able to be amplified from any cDNA sample, possibly due to the less than optimal temperature difference (of ~10 ^{o}C) between the forward and reverse pair; this was an unfortunate necessity as the 5' region of the cDNA sequence for *ROCK2* was extremely GC-rich, raising the required temperature for denaturation. This may be overcome in later work by designing a paired reverse primer in a similarly GC-rich region, although there are well reported problems with GC-rich primer amplification also (Frey et al., 2008). Another possible strategy may be to attempt amplification using the genomic DNA sequence available in the reference genome, now that the sequence of the ovine cDNA for the 3' region from this target area can be identified.

Fragment F was not able to be sequenced in the affected individual, despite an apparent product observed from electrophoresis. While this was unfortunate, if the disease is caused by a deletion in 27', the *ROCK2m* transcript may not be present in the affected individual, and the presence of a homozygous mutant transcript in muscle not normally affected by the pathology is perhaps unlikely. The muscle wasting in the affected individual prevented the use of a 'normally affected' muscle for RNA extraction, though this could potentially be overcome by excising soleus tissue from a younger affected individual (e.g. RED 1 – 4) as this muscle wasting is progressive over

the life of the sheep. The more important finding is to determine whether there is a heterozygous mutation in the carrier sequence, as a 'normally affected' muscle in the carrier would express *ROCK2m*, but possess a heterozygous mutation.

Fragment H was not able to be amplified from the affected gracilus cDNA. A possible cause for this failure is that the affected gracilus cDNA is not amplifying for fragment H as there is a deletion under one or both of the primer locations, and this could quite plausibly be the genetic cause of the disease. As an affected individual is predicted to be homozygous for the causative mutation, whereas a carrier only heterozygous, one would expect the absence of PCR product for the "affected" cDNA but the presence of a product from a "carrier" cDNA. Sequencing for fragment H was unsuccessful for both the carrier gracilus and the carrier soleus, despite relatively clear banding for each (Fig. 3.1.). This may be the result of the additional unclear bands being indicative of both transcripts being present, the major isoform of *ROCK2*, and *ROCK2m*, thus interfering with the sequencing results. A possible solution may be to excise the larger product bands from a gel, purify, and sequence this, or to clone a mixed pool of products and sequence individual clones, as each clone would only contain one or other fragment. Amplification of these fragments from cDNA will be further attempted in continued studies.

Fragments F and H would encompass the predicted location of 27' in sheep, spanning exons 24 – 32 and exons 24 – 3' UTR, respectively. Unfortunately the PCR products for segement H did not sequence well and due to the limitations of time these experiments could not be repeated to include in this thesis work. However, the successful sequencing of fragment F using the forward primer from the carrier soleus and gracilus in this work demonstrated the presence in the OCPMD sheep of a region analogous to 27', this being of 171 bp, distinctly different from the reference sequence and that of both reverse strands (which were in agreement with the reference) Further, the bases towards the end of the reverse sequencing showed agreement with that of the putative 27' at an equivalent position as shown in the forward strand sequence. Additionally, a drop in the intensity of chromatogram peaks at approximately the start of this region (Fig. 3.2.) indicated that the region contained two products. This identification has enabled the design of primers specific to 27', which will allow the sequencing of this region in not only the carriers but also, we expect, in the affected individuals of the OCPMD flock. At this stage we have successfully sequenced ~90% of the *ROCK2* cDNA in the carrier, and ~80% of the sequence in the affected.

Discovering a variant within 27' which is homozygous within the affected and heterozygous within the carrier will provide very strong evidence for the involvement of *ROCK2m* with this unique disease pathology. Rho kinases like *ROCK2* have been reported to be potential therapeutic targets in the treatment of cardiovascular disease (Surma et al., 2011), but *ROCK2* has never before been implicated in causing any disease, let alone one of skeletal muscle. If proven through additional work as being the disease gene this will provide the possibility of widening the genetic screening net, and assisting in the characterisation of human disease which has not yet had its genetic cause determined, such as the phenotypically similar LGMD2G (Paim et al., 2013). If a rare human disease like LGMD2G is able to be linked to this gene, the ovine muscular dystrophy model will additionally provide a valuable opportunity to study the pathogenesis in a way not possible in a human population.

This project has identified the most likely location for a variant causative of the OCPMD pathology. Primers enabling closer interrogation of sheep 27' have been designed based on the sequencing results of this project and this sequencing will be subsequently undertaken. If a potentially causative variant is identified, replication will be undertaken using the already collected biological samples from a large number of individuals of the flock. The strategies outlined above will be used to amplify and sequence the remaining fragments which were unsuccessful in this work. In the absence of a heterozygous variant being identified in the fully sequenced cDNA of *ROCK2*, the genomic DNA of *ROCK2* may be sequenced using either traditional (Sanger) or next-generation techniques. A mutation in the promoter region upstream of *ROCK2* may also affect *ROCK2* or *ROCK2m* expression, and this could be investigated using quantitative RT-PCR.

In addition, the authors of the 2007 paper (Pelosi et al., 2007) providing first evidence of the existence of 27', and *ROCK2m*, have been contacted in search of the novel antibodies they developed for the isoform. If this is able to be secured it will provide the ability to carry out histochemical analysis to examine differential expression between carriers and affected sheep of the flock. Even in the absence of this antibody, western blotting using a generic *ROCK2* antibody may provide a valuable avenue of approach.

The ratio of the transcripts of *ROCK2* to *ROCK2m* in skeletal muscles was reported to vary between muscles in mice (Pelosi et al., 2007). By quantitatively amplifying the

ROCK2m fragment in different muscles we may be able to determine whether the muscles exhibiting pathology in the OCPMD flock are those which have a greater proportion of *ROCK2m*. The specificity of *ROCK2m* to type I fibres in comparison to type II has not been previously examined, and the combination of *ROCK2m*-immunostaining along with myosin heavy chain antibodies to detect different myofiber types may provide additional evidence for or against this as a disease gene candidate.

If all of these investigations return negative evidence for *ROCK2* involvement in the OCPMD pathology, we may investigate the next most significant candidate from our bioinformatics analyses, or carry out a higher coverage sequencing of an affected and carrier individual.

CONCLUSIONS

This ambitious, multi-disciplinary project combining both *in silico* and molecular techniques has identified a prime candidate gene for a causative genetic variant in the OCPMD flock. It has additionally provided the first evidence of *ROCK2m* in sheep and laid the groundwork for further investigations into *ROCK2* and the OCPMD pathology.

5. Bibliography

1974. Federal Nonnuclear Energy Research and Development Act of 1974. *In:* Senate and House of Representatives of the United States of America in Congress Assembled (ed.).

1990. Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years. *In:* United States Department of Energy, United States Department of Health and Human Services (ed.).

2013. Picard-Tools. v. 1.91 ed. http://sourceforge.net/projects/picard/files/picard-tools/1.91/picard-tools-1.91.zip/download

Abecasis, G., Cherny, S., Cookson, W. & Cardon, L. 2002. Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics,* 30**,** 97 - 101.

Abresch, R. T., Walsh, S. A. & Wineinger, M. A. 1998. Animal models of neuromuscular diseases: pathophysiology and implications for rehabilitation. *Physical medicine and rehabilitation clinics of North America,* 9**,** 285-299.

Access Economics 2007. The Cost of Muscular Dystrophy. Muscular Dystrophy Association.

Alkan, C., Coe, B. P. & Eichler, E. E. 2011. Genome structural variation discovery and genotyping. *Nature Reviews: Genetics,* 12**,** 363-376.

Allamand, V. & Campbell, K. P. 2000. Animal models for muscular dystrophy: valuable tools for the development of therapies. *Human Molecular Genetics,* 9**,** 2459-2467.

Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. 2001. Genomewide scans of complex human diseases: true linkage is hard to find. *American Journal of Human Genetics,* 69**,** 936-950.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-410.

Anderson, S., Bankier, A. T., Barrell, B. G., Bruijn, M. H. L. D., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. 1981. Sequence and organization of the human mitochondrial genome. *Nature,* 290**,** 457-465.

Archibald, A. L., Cockett, N. E., Dalrymple, B. P., Faraut, T., Kijas, J. W., Maddox, J. F., Mcewan, J. C., Hutton Oddy, V., Raadsma, H. W., Wade, C., Wang, J., Wang, W. & Xun, X. 2010. The sheep genome reference sequence: a work in progress. *Animal Genetics,* 41**,** 449-453.

Ashizawa, T. & Sarkar, P. 2011. Myotonic dystrophy types 1 and 2. *Handbook of Clinical Neurology,* 101**,** 193-237.

Barbujani, G., Russo, A., Danieli, G. A., Spiegler, A. W. J., Borkowska, J. & Petrusewicz, I. H. 1990. Segregation analysis of 1885 DMD families: significant

departure from the expected proportion of sporadic cases. *Human Genetics,* 84**,** 522-526.

Bartlett, J. M. S. & Stirling, D. 2003. A Short History of the Polymerase Chain Reaction. *In:* BARTLETT, J. M. S. & STIRLING, D. (eds.) *PCR Protocols.* Humana Press.

Becker, J., Kleinsmith, L., Hardin, J. & Bertoni, G. 2008. *The World of the Cell,* Pearson Education Inc.

Beenakker, E. C., Fock, J. M., Van Tol, M. J. & Et Al. 2005. Intermittent prednisone therapy in duchenne muscular dystrophy: A randomized controlled trial. *Archives of Neurology,* 62**,** 128-132.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. 2000. GenBank. *Nucleic Acids Research,* 28**,** 15-18.

Bertini, E., D'amico, A., Gualandi, F. & Petrini, S. 2011. Congenital Muscular Dystrophies: A Brief Review. *Seminars in Pediatric Neurology,* 18**,** 277-288.

Blake, D. J., Weir, A., Newey, S. E. & Davies, K. E. 2002. Function and genetics of dystrophin and dystrophin-related proteins in muscle. *Physiology Review,* 82**,** 291-329.

Boguski, M. S. 1998. Bioinformatics - a new era. *Trends Guide to Bioinformatics***,** 1-3.

Botstein, D. & Risch, N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet,* 33**,** 228-237.

Bresolin, N., Castelli, E., Comi, G. P., Felisari, G., Bardoni, A., Perani, D., Grassi, F., Turconi, A., Mazzucchelli, F., Gallotti, D., Moggio, M., Prelle, A., Ausenda, C., Fazio, G. & Scarlato, G. 1994. Cognitive impairment in Duchenne muscular dystrophy. *Neuromuscular Disorders,* 4**,** 359-369.

Brook, J. D., Mccurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J.-P., Hudson, T., Sohn, R., Zemelman, B., Snell, R. G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P. S., Shaw, D. J. & Housman, D. E. 1992. Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 32 end of a transcript encoding a protein kinase family member. *Cell,* 68**,** 799-808.

Campbell, K. P. 1995. Three muscular dystrophies: Loss of cytoskeleton-extracellular matrix linkage. *Cell,* 80**,** 675-679.

Cartegni, L., Chew, S. L. & Krainer, A. R. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet,* 3**,** 285-298.

Chao, K.-M. 2012. *Introduction to SNP and Haplotype Analysis* National Taiwan University.

Check Hayden, E. 2009. Genome sequencing: the third generation. *Nature,* 457**,** 768-9.

Chen, L., Tovar-Corona, J. M. & Urrutia, A. O. 2012. Alternative Splicing: A Potential Source of Functional Innovation in the Eukaryotic Genome. *International Journal of Evolutionary Biology,* 2012**,** 10.

Clark, D. P. 2005. *Molecular Biology,* Elsevier Academic Press.

Cohen, H. J., Molnar, G. E. & Taft, L. T. 1968. The Genetic Relationship of Progressive Muscular Dystrophy (Duchenne Type) and Mental Retardation. *Developmental Medicine & Child Neurology,* 10**,** 754-765.

Collins, F. S., Morgan, M. & Patrinos, A. 2003. The Human Genome Project: Lessons from Large-Scale Biology. *Science,* 300**,** 286-290.

Conway, T. & K, G. 2003. Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. *Molecular Microbiology,* 47**,** 879-889.

Corporation, C. 2013. Available: www.codoncode.com.

Crotti, L. B. & Horowitz, D. S. 2009. Exon sequences at the splice junctions affect splicing fidelity and alternative splicing. *Proceedings of the National Academy of Sciences, U.S.A.,* 106**,** 18954-18959.

Cui, Y., Li, G., Li, S. & Wu, R. 2010. Designs for Linkage Analysis and Association Studies of Complex Diseases. *In:* BANG, H., ZHOU, X. K., EPPS, H. L. & MAZUMDAR, M. (eds.) *Statistical Methods in Molecular Biology.* Humana Press.

Cunningham, J. G. K., B. G. 2007. *Veterinary Physiology,* Saunders.

Danièle, N., Richard, I. & Bartoli, M. 2007. Ins and outs of therapy in limb girdle muscular dystrophies. *The International Journal of Biochemistry & Cell Biology,* 39**,** 1608-1624.

Davies, K. E. & Nowak, K. J. 2006. Molecular mechanisms of muscular dystrophies: old and new players. *Nature Reviews Molecular Cell Biology,* 7**,** 762-773.

Dayhoff, M. O. 1965. Computer aids to protein sequence determination. *Journal of Theoretical Biology,* 8**,** 97-112.

Dayhoff, M. O. L., R. S. 1962. Compoprotein: A computer program to aid primary protein structure determination. *Procedures of the Fall Joint Computer Conference,* 22**,** 262-274.

Den Dunnen, J. T., Grootscholten, R. M., Bakker, E., Blonden, L. a. J., Ginjaar, H. B., Wapenaar, M. C., Van Paassen, H. M. B., Van Broeckhoven, C., Pearson, R. L. & Van Ommen, G. J. B. 1989. Topography of the Duchenne Muscular Dystrophy (DMD) Gene: FIGE and cDNA Analysis of 194 Cases Reveals 115 Deletions and 13 Duplications. *American Journal of Human Genetics,* 45**,** 835-847.

Dent, A. G., Richards, R. B. & Nairn, M. E. 1979. Congenital Progressive Ovine Muscular Dystrophy in Western Australia. *Australian Veterinary Journal,* 55**,** 297-297.

Devlin, B. & Risch, N. 1995. A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics,* 29**,** 311-322.

Dhand, R. 2006. The 'finished' landscape. *Nature,* S1 (2006)**,** 7.

Di Maio, L., Boiano, S., Squitieri, F., Napolitano, G., Cocozza, S., Campanella, G. & Battistuzzi, G. 1992. Genetic linkage analysis and presymptomatic testing in Huntington's disease. First report in Italy. *Acta Neurol (Napoli),* 14**,** 524-9.

Doherty, J. T., Lenhart, K. C., Cameron, M. V., Mack, C. P., Conlon, F. L. & Taylor, J. M. 2011. Skeletal Muscle Differentiation and Fusion Are Regulated by the BAR-containing Rho-GTPase-activating Protein (Rho-GAP), GRAF1. *Journal of Biological Chemistry,* 286**,** 25903-25921.

Dolled-Filhart, M. P., Lee, M., Ou-Yang, C.-W., Haraksingh, R. R. & Lin, J. C.-H. 2013. Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing. *The Scientific World Journal,* 2013**,** 10.

Dubowitz, V. 1965. Intellectual Impairement in Muscular Dystrophy. *Archives of Disease in Childhood,* 40**,** 296 - 301.

Embl-Ebi 2013. EMBLE-EBI Database of Complete Virus Genomes. Page generated 4:01pm Mon 22-JUL-2013 ed.: Wellcome Trust Genome Campus, Hinxton, Cambridgeshire.

Emery, A. E. H. 1991. Population frequencies of inherited neuromuscular diseases - A world survey. *Neuromuscular Disorders,* 1**,** 19-29.

Emery, A. E. H. 2002. The muscular dystrophies. *The Lancet,* 359**,** 687-695.

Fairclough, R. J., Wood, M. J. & Davies, K. E. 2013. Therapy for Duchenne muscular dystrophy: renewed optimism from genetic approaches. *Nat Rev Genet,* 14**,** 373-378.

Fan, B., Du, Z.-Q., Gorbach, D. M. & Rothschild, M. F. 2010. Development and Application of High-density SNP Arrays in Genomic Studies of Domestic Animals. *Asian-Australian Journal of Animal Science,* 23**,** 833-847.

Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., Radlwimmer, B. & Stange, E. F. 2006. A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the Colon. *American Journal of Human Genetics,* 79**,** 439-448.

Fitch, W. M. 1966. An improved method of testing for evolutionary homology. *Journal of Molecular Biology,* 16**,** 9-16.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., Mckenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocyne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., Mcdonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. 1995. Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. *Nature,* 269**,** 496-498.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A. K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., Mclaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J. P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A. & Searle, S. M. J. 2013. Ensembl 2013. *Nucleic Acids Research,* 41**,** D48-D55.

Frey, U. H., Bachmann, H. S., Peters, J., Siffert, W. 2008 PCR-amplification of GC-rich regions: 'slowdown PCR'. *Nature Protocols* 3(8), 1312-7

Fossett, N. G., Arbour-Reily, P., Kilroy, G., Mcdaniel, M., Mahmoud, J., Tucker, A. B., Chang, S. H. & Lee, W. R. 1990. Analysis of ENU-induced mutations at the Adh locus in Drosophila melanogaster. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* 231**,** 73-85.

Francomano, C. A. & Kazazian, H. H. 1986. DNA Analysis In Genetic Disorders. *Annual Review of Medicine,* 37**,** 377-95.

Frazer, K. A. 2012. Decoding the human genome. *Genome Research,* 22**,** 1599-1601.

Fukuda, Y., Nakahara, Y., Date, H., Takahashi, Y., Goto, J., Miyashita, A., Kuwano, R., Adachi, H., Nakamura, E. & Tsuji, S. 2009. SNP HiTLink: a high-throughput linkage analysis system employing dense SNP data. *BMC Bioinformatics,* 10**,** 121.

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D. & Province, M. A. 2010. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology,* 34**,** 100-105.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. 1996. Life with 6000 Genes. *Science,* 274**,** 546-567.

Gonzalez-Galarza, F. F., Lawless, C., Hubbard, S. J., Fan, J., Bessant, C., Hermjakob, H. & Jones, A. R. 2012. A Critical Appraisal of Techniques, Software Packages, and Standards for Quantitative Proteomic Analysis. *OMICS,* 16**,** 431-442.

Gozal, D. 2000. Pulmonary manifestations of neuromuscular disease with special reference to Duchenne muscular dystrophy and spinal muscular atrophy. *Pediatric Pulmonology,* 29**,** 141-150.

Hagen, J. B. 2000. The origins of bioinformatics. *Nature Reviews Genetics,* 1**,** 231+.

Halperin, E., Kimmel, G. & Shamir, R. 2005. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics,* 21**,** i195-i203.

Hardy, J. & Singleton, A. 2009. Genomewide Association Studies and Human Disease. *New England Journal of Medicine,* 360**,** 1759-1768.

Hartl, D. L. & Clark, A. G. 1997. *Principles of Population Genetics,* Sunderland: Sinauer Associates.

Hawkins, R. D., Hon, G. C. & Ren, B. 2010. Next-generation genomics: an integrative approach. *Nature Reviews Genetics,* 11**,** 476-486.

Hicks, D., Sarkozy, A., Muelas, N., Köehler, K., Huebner, A., Hudson, G., Chinnery, P. F., Barresi, R., Eagle, M., Polvikoski, T., Bailey, G., Miller, J., Radunovic, A., Hughes, P. J., Roberts, R., Krause, S., Walter, M. C., Laval, S. H., Straub, V., Lochmüller, H. & Bushby, K. 2011. A founder mutation in Anoctamin 5 is a major cause of limb girdle muscular dystrophy. *Brain,* 134**,** 171-182.

Hill, W. G. & Robertson, A. 1968. The Effects of Inbreeding at Loci with Heterozygote Advantage. *Genetics,* 60**,** 615-628.

Horner, D. S., Pavesi, G., Castrignanò, T., De Meo, P. D., Liuni, S., Sammeth, M., Picardi, E. & Pesole, G. 2010. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics,* 11**,** 181-197.

Hunter, D. J. 2005. Gene-environment interactions in human diseases. *Nature Reviews Genetics,* 6**,** 287-298.

Ianello-Giassetti, M., Fernanda-Sevciuc, M., D'Ávila-Assumpção, M. E. & Antônio-Visintin, J. 2013. Genetic Engineering and Cloning: Focus on Animal Biotechnology. *Genetic Engineering,* InTech

Ilkovski, B., Cooper, S. T., Nowak, K., Ryan, M. M., Yang, N., Schnell, C., Durling, H. J., Roddick, L. G., Wilkinson, I., Kornberg, A. J., Collins, K. J., Wallace, G., Gunning, P., Hardeman, E. C., Laing, N. G. & North, K. N. 2001. Nemaline Myopathy Caused by Mutations in the Muscle ±-Skeletal-Actin Gene. *American Journal of Human Genetics,* 68**,** 1333-1343.

International-Human-Genome-Sequencing-Consortium 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-945.

Kakulas, B. A., Adams, R. D. (ed.) 1985. *Diseases of muscle: pathological foundations of clinical myology,* Sydney: Harper & Row.

Kanehisa, M. & Bork, P. 2003. Bioinformatics in the post-sequence era. *Nature Genetics,* 33.

Kanehisa, M. & Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research,* 28**,** 27-30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research,* 40**,** D109-D114.

Karagan, N. J. 1979. Intellectual functioning in Duchenne muscular dystrophy: A review. *Psychological Bulletin,* 86**,** 250-259.

Kerem, B.-S. Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M., Tsui, L.-C. 1989. Identification of the Cystic Fibrosis Gene: Genetic Analysis. *Science,* 245.

Kerem, E., Corey, M., Kerem, B.-S., Rommens, J., Markiewicz, D., Levison, H., Tsui, L.-C. & Durie, P. 1990. The Relation between Genotype and Phenotype in Cystic Fibrosis — Analysis of the Most Common Mutation (ΔF508). *New England Journal of Medicine,* 323**,** 1517-1522.

Klesert, T. R., Otten, A. D., Bird, T. D. & Tapscott, S. J. 1997. Trinucleotide repeat expansion at the myotonic dystrophy locus reduces expression of *DMAHP*. *Nature Genetics,* 16**,** 402-406.

Kondo-Lida, E., Kobayashi, K., Watanabe, M., Sasaki, J., Kumagai, T., Koide, H., Saito, K., Osawa, M., Nakamura, Y. & Toda, T. 1999. Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD). *Human Molecular Genetics,* 8**,** 2303-2309.

Korbie, D. J. & Mattick, J. S. 2008. Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nature Protocols,* 3**,** 1452-1456.

Kumar, S. & Dudley, J. 2007. Bioinformatics software for biologists in the genomics era. *Bioinformatics,* 23**,** 1713-1717.

Kwiatkowska, J. & Slomski, R. 1992. [DMD gene--the largest human gene]. *Postepy Biochemii,* 38**,** 49-55.

Laframboise, T. 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research,* 37**,** 4181-4193.

Laing, N. G. 2012. Genetics of neuromuscular disorders. *Critical Reviews in Clinical Laboratory Sciences,* 49**,** 33-48.

Laval, S. H. & Bushby, K. M. D. 2004. Limb-girdle muscular dystrophies – from genetics to molecular pathology. *Neuropathology and Applied Neurobiology,* 30**,** 91-105.

Leibowitz, D. & Dubowitz, V. 1981. Intellect and Behaviour in Duchenne Muscular Dystrophy. *Developmental Medicine & Child Neurology,* 23**,** 577-590.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., Macdonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., Mcintosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y.-H., Frazier, M. E., Scherer, S. W., Strausberg, R. L. & Venter, J. C. 2007. The Diploid Genome Sequence of an Individual Human. *PLoS Biol,* 5**,** e254.

Li, H. D., R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics,* 26**,** 589 - 595.

Li, M., Boehnke, M. & Abecasis, G. R. 2005. Joint Modeling of Linkage and Association: Identifying SNPs Responsible for a Linkage Signal. *The American Journal of Human Genetics,* 76**,** 934-949.

Lipman, D. J. & Pearson, W. R. 1985. Rapid and Sensitive Protein Similarity Searches. *Science,* 227**,** 1435-1441.

Longman, C. 2006. Myotonic dystrophy. *Journal of the Royal College of Physicians of Edinburgh,* 36**,** 51-55.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R. 2005. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters,* 579**,** 1900-1903.

Luscombe, N. M., Greenbaum, D. & Gerstein, M. 2001. What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics* [Online].

Lutton, J. D., Winston, R. & Rodman, T. C. 2003. Multiple Sclerosis: Etiological Mechanisms and Future Directions. *Experimental Biology and Medicine,* 229**,** 12-20.

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F. & Brandi, M. L. 2010. Bioinformatics for Next Generation Sequencing Data. *Genes,* 1**,** 294-307.

Mardis, E. 2010. The $1,000 genome, the $100,000 analysis? *Genome Medicine,* 2**,** 84.

Marston, A. L. & Amon, A. 2004. Meiosis: cell-cycle controls shuffle and deal. *Nature Reviews Molecular Cell Biology,* 5**,** 983-997.

Maskos, U. & Southern, E. M. 1992. Oligonucleotide hybridisations on glass supports: a novel linker for oligonucleotide synthesis and hybridisation properties of oligonucleotides synthesised in situ. *Nucleic Acids Research,* 20**,** 1679-1684.

McDonald, C. M. 2002. Physical Activity, Health Impairments, and Disability in Neuromuscular Disease. *American Journal of Physical Medicine & Rehabilitation,* 81**,** S108-S120.

McGavin, M. D. & Baynes, I. D. 1969. A Congenital Progressive Ovine Muscular Dystrophy *Veterinary Pathology,* 6**,** 513-524.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & Depristo, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* 20**,** 1297 - 1303.

McKusick, V. A. 2013. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD).

Muscular Dystrophy Association. 2013. *How Are Neuromuscular Diseases Treated?* [Online]. Muscular Dystrophy Association. Available: http://mda.org/publications/teachers-guide/how-are-NMDs-treated [Accessed 27 July 2013 2013].

Meltzer, P. S. 2001. Spotting the target: microarrays for disease gene discovery. *Current Opinion in Genetics & Development,* 11**,** 258-263.

Mercuri, E., Pichiecchio, A., Allsop, J., Messina, S., Pane, M. & Muntoni, F. 2007. Muscle MRI in inherited neuromuscular disorders: Past, present, and future. *Journal of Magnetic Resonance Imaging,* 25**,** 433-440.

Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nature Reviews Genetics,* 11**,** 31-46.

Miteff, F., Potter, H. C., Allen, J., Teoh, H., Roxburgh, R. & Hutchinson, D. O. 2011. Clinical and laboratory features of patients with myophosphorylase deficiency (McArdle disease). *Journal of Clinical Neuroscience: Official Journal of the Neurosurgical Society of Australasia,* 18**,** 1055-1058.

Moore, J. H., Asselbergs, F. W. & Williams, S. M. 2010. Bioinformatics challenges for genome-wide association studies. *Bioinformatics,* 26**,** 445-455.

Morales, F., Couto, J. M., Higham, C. F., Hogg, G., Cuenca, P., Braida, C., Wilson, R. H., Adam, B., Del Valle, G., Brian, R., Sittenfeld, M., Ashizawa, T., Wilcox, A., Wilcox, D. E. & Monckton, D. G. 2012. Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Human Molecular Genetics,* 21**,** 3558-3567.

Mullard, A. 2007. Silent mutations turn up the volume. *Nature Reviews Molecular Cell Biology,* 8**,** 98-98.

Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G. & Stahl, F. 2013. Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology,* 24**,** 22-30.

Ng, R., Banks, G. B., Hall, J. K., Muir, L. A., Ramos, J. N., Wicki, J., Odom, G. L., Konieczny, P., Seto, J., Chamberlain, J. R. & Chamberlain, J. S. 2012. Animal Models of Muscular Dystrophy. *In:* CONN, P. M. (ed.) *Progress in Molecular Biology and Translational Science.* Academic Press.

Nielsen, R., Paul, J. S., Albrechtson, A. & Song, Y. S. 2011. Genotyping and SNP calling from next-generation sequencing data. *Nature Reviews,* 12**,** 443-451.

Nigro, V. 2003. Molecular bases of autosomal recessive limb-girdle muscular dystrophies. *Acta Myologica,* 22**,** 35-42.

North, K. N., Laing, N. G. & Wallgren-Pettersson, C. 1997. Nemaline myopathy: current concepts. The ENMC International Consortium and Nemaline Myopathy. *Journal of Medical Genetics,* 34**,** 705-713.

Nowak, K. J. & Davies, K. E. 2004. Duchenne muscular dystrophy and dystrophin: pathogenesis and opportunities for treatment. *EMBO Rep,* 5**,** 872-876.

Oliver, S. G., Van Der Aart, Q. J., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J. P., Benit, P. & Et Al. 1992. The complete DNA sequence of yeast chromosome III. *Nature,* 357**,** 38-46.

Ong, C. E., Pan, Y., Mak, J. W. & Ismail, R. 2013. *In vitro* approaches to investigate cytochrome P450 activities: update on current status and their applicability. *Expert Opinion on Drug Metabolism & Toxicology,* 0**,** 1-17.

Paim, J. F., Cotta, A., Vargas, A. P., Navarro, M. M., Valicek, J., Carvalho, E., Da-Cunha-Junior, A. L., Plentz, E., Braz, S. V., Takata, R., Almeida, C. F. & Vainzof, M. 2013. Muscle Phenotypic Variability in Limb Girdle Muscular Dystrophy 2 G. *Journal of Molecular Neuroscience,* 50**,** 339-344.

Pedrosa, D. J. & Timmermann, L. 2013. Review: management of Parkinson's disease. *Journal of Neuropsychiatric Disease Treatment,* 9**,** 321-40.

Pelosi, M., Marampon, F., Zani, B. M., Prudente, S., Perlas, E., Caputo, V., Cianetti, L., Berno, V., Narumiya, S., Kang, S. W., Musaro, A. & Rosenthal, N. 2007. ROCK2 and Its Alternatively Spliced Isoform ROCK2m Positively Control The Maturation of the Myogenic Program. *Molecular and Cellular Biology,* 27**,** 6163-6176.

Perel, P., Roberts, I., Sena, E., Wheble, P., Briscoe, C., Sandercock, P., Macleod, M., Mignini, L. E., Jayaram, P. & Khan, K. S. 2007. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ,* 334**,** 197.

Pierce, B. A. 2010. *Genetics: A Conceptual Approach,* New York, W.H. Freeman and Company.

Prather, R. S., Lorson, M., Ross, J. W., Whyte, J. J. & Walters, E. 2013. Genetically Engineered Pig Models for Human Diseases. *Annual Review of Animal Biosciences,* 1**,** 203-219.

Purcell, S. 2013. PLINK v.1.07. http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J. & Sham, P. C. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics,* 81**,** 559 - 575.

Qiagen 2010. RNeasy Fibrous Tissue Handbook.

Ranganatha, N., Kuppast, I. J. 2012. A review on alternatives to animal testing methods in drug development. *International Journal of Pharmazy and Pharmaceutical Sciences,* 4**,** 28-32.

Ravenscroft, G., Miyatake, S., Lehtokari, V.-L., Todd, Emily j., Vornanen, P., Yau, Kyle s., Hayashi, Yukiko k., Miyake, N., Tsurusaki, Y., Doi, H., Saitsu, H., Osaka, H., Yamashita, S., Ohya, T., Sakamoto, Y., Koshimizu, E., Imamura, S., Yamashita, M., Ogata, K., Shiina, M., Bryson-Richardson, Robert j., Vaz, R., Ceyhan, O., Brownstein, Catherine a., Swanson, Lindsay c., Monnot, S., Romero, Norma b., Amthor, H., Kresoje, N., Sivadorai, P., Kiraly-Borri, C., Haliloglu, G., Talim, B., Orhan, D., Kale, G., Charles, Adrian k., Fabian, Victoria a., Davis, Mark r., Lammens, M., Sewry, Caroline a., Manzur, A., Muntoni, F., Clarke, Nigel f., North, Kathryn n., Bertini, E., Nevo, Y., Willichowski, E., Silberg, Inger e., Topaloglu, H., Beggs, Alan h., Allcock, Richard j. N., Nishino, I., Wallgren-Pettersson, C., Matsumoto, N. & Laing, Nigel g. 2013. Mutations in KLHL40 Are a Frequent Cause of Severe Autosomal-Recessive Nemaline Myopathy. *American Journal of Human Genetics,* 93**,** 6-18.

Richards, R. B., Lewer, R. P., Passmore, I. K. & Mcquade, N. C. 1988b. Ovine congenital progressive muscular dystrophy: mode of inheritance. *Australian Veterinary Journal,* 65**,** 93-94.

Richards, R. B. & Passmore, I. K. 1989. Ultrastructural changes in skeletal muscle in ovine muscular dystrophy. *Acta Neuropathologica,* 79**,** 168-175.

Richards, R. B., Passmore, I. K. & Dempsey, E. F. 1988a. Skeletal muscle pathology in ovine congenital progressive muscular dystrophy. *Acta Neuropathol,* 77**,** 161-167.

Richards, R. B., Passmore, J. K., Bretag, A. H., Kakulas, B. A. & Mcquade, N. C. 1986. Ovine congenital progressive muscular dystrophy: clinical syndrome and distribution of lesions. *Australian Veterinary Journal,* 63**,** 396-401.

Rochester, D. F. & Esau, S. A. 1994. Assessment of ventilatory function in patients with neuromuscular disease. *Clinics in Chest Medicine,* 14**,** 751-63.

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Meur, N. L., Laquerriere, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., Dubas, F., Frebourg, T. & Campion, D. 2006. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genetics,* 38**,** 24-26.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M. & Smith, M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature,* 265**,** 687-95.

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. 1982. Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology,* 162**,** 729-773.

Sboner, A., Mu, X., Greenbaum, D., Auerbach, R. & Gerstein, M. 2011. The real cost of sequencing: higher than you think! *Genome Biology,* 12**,** 125.

Schadt, E. E., Turner, S. & Kasarskis, A. 2010. A window into third-generation sequencing. *Human Molecular Genetics,* 19**,** R227-40.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A. & Wigler, M. 2004. Large-Scale Copy Number Polymorphism in the Human Genome. *Science,* 305**,** 525-528.

Sellick, G. S., Longman, C., Tolmie, J., Newbury-Ecob, R., Geenhalgh, L., Hughes, S., Whiteford, M., Garrett, C. & Houlston, R. S. 2004. Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays. *Nucleic Acids Research,* 32**,** e164.

Service, R. F. 2006. The Race for the $1000 Genome. *Science,* 311**,** 1544-1546.

Shendure, J. & Ji, H. 2008. Next-generation DNA sequencing. *Nature Biotechnology,* 26**,** 1135-1145.

Sparknotes-Editors. 2013. *SparkNote on Post-Transcriptional RNA Processing* [Online]. SparkNotes LLC. [Accessed 29 June 2013 2013].

Stajich, J. E. & Lapp, H. 2006. Open source tools and toolkits for bioinformatics: significance, and where are we? *Briefings in Bioinformatics,* 7**,** 287-296.

Surma, M., Wei, L. & Shi, J. 2011. Rho kinase as a therapeutic target in cardiovascular disease. *Future Cardiology,* 7**,** 657-671.

Syvanen, A.-C. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics,* 2**,** 930-942.

Tempest, H. G. 2011. Meiotic recombination errors, the origin of sperm aneuploidy and clinical recommendations. *Systems Biology in Reproductive Medicine,* 57**,** 93-101.

Tenesa, A. & Haley, C. S. 2013. The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics,* 14**,** 139-149.

Teufel, A., Krupp, M., Weinmann, A. & Galle, P. R. 2006. Current bioinformatics tools in genomic biomedical research (Review). *International Journal of Molecular Medicine,* 17**,** 967-973.

Tews, D. S. 2002. Apoptosis and muscle fibre loss in neuromuscular disorders. *Neuromuscular Disorders,* 12**,** 613-622.

The International Hapmap Consortium 2005. A haplotype map of the human genome. *Nature,* 437**,** 1299-1320.

The International Multiple Sclerosis Genetics Consortium 2010. Evidence for Polygenic Susceptibility to Multiple Sclerosis—The Shape of Things to Come. *The American Journal of Human Genetics,* 86**,** 621-625.

Thyagarajan, T., Totey, S., Danton, M. J. S. & Kulkarni, A. B. 2003. Genetically Altered Mouse Models: the Good, the Bad, and the Ugly. *Critical Reviews in Oral Biology & Medicine,* 14**,** 154-174.

Turner, C. & Hilton-Jones, D. 2010. The myotonic dystrophies: diagnosis and management. *Journal of Neurology, Neurosurgery & Psychiatry,* 81**,** 358-367.

Urtasun, M., Sáenz, A., Roudaut, C., Poza, J. J., Urtizberea, J. A., Cobo, A. M., Richard, I., García Bragado, F., Leturcq, F., Kaplan, J. C., Martí Massó, J. F., Beckmann, J. S. & López De Munain, A. 1998. Limb-girdle muscular dystrophy in Guipúzcoa (Basque Country, Spain). *Brain,* 121**,** 1735-1747.

Vallee, M., Robert, C., Methot, S., Palin, M.-F. & Sirard, M.-A. 2006. Cross-species hybridizations on a multi-species cDNA microarray to identify evolutionarily conserved genes expressed in oocytes. *BMC Genomics,* 7**,** 113.

Van Der Kooi, A. J., Barth, P. G., Busch, H. F. M., De Haan, R., Ginjaar, H. B., Van Essen, A. J., Van Hooff, L. J. M. A., Höweler, C. J., Jennekens, F. G. I., Jongen, P., Oosterhuis, H. J. G. H., Padberg, G. W. a. M., Spaans, F., Wintzen, A. R., Wokke, J. H. J., Bakker, E., Van Ommen, G. J. B., Bolhuis, P. A. & De Visser, M. 1996. The clinical spectrum of limb girdle muscular dystrophy A survey in the Netherlands. *Brain,* 119**,** 1471-1480.

Van Der Worp, H. B., Howells, D. W., Sena, E. S., Porritt, M. J., Rewell, S., O'collins, V. & Macleod, M. R. 2010. Can Animal Models of Disease Reliably Inform Human Studies? *PLoS Med,* 7**,** e1000245.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., Mckusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A.,

Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., et al. 2001. The Sequence of the Human Genome. *Science,* 291**,** 1304-1351.

Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O. & Hunkapiller, M. 1998. Shotgun Sequencing of the Human Genome. *Science,* 280**,** 1540-1542.

Visscher, P. M., Brown, M. A., Mccarthy, M. I. & Yang, J. 2012. Five Years of GWAS Discovery. *The American Journal of Human Genetics,* 90**,** 7-24.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M. & Lewis, S. E. 2009. Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation. *PLoS Biol,* 7**,** e1000247.

Waterston, R. H., Lander, E. S. & Sulston, J. E. 2002. On the sequencing of the human genome. *Proceedings of the National Academy of Sciences,* 99**,** 3712-3716.

Weiner, H. L. 2004. Immunosuppressive treatment in multiple sclerosis. *Journal of the Neurological Sciences,* 223**,** 1-11.

Wellcome Trust Case Control Consortium 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-678.

Wetterstrand, K. A. 2013. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)* [Online]. Available: www.genome.gov/sequencingcosts [Accessed 24 Jul 2013 2013].

Wigginton, J. E. & Abecasis, G. R. 2005. PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics,* 21**,** 3445-3447.

Woese, C. R. & Fox, G. E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences,* 74**,** 5088-5090.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. & Madden, T. 2012. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics,* 13**,** 134.

Zatkova, A., Messiaen, L., Vandenbroucke, I., Wieser, R., Fonatsch, C., Krainer, A. R. & Wimmer, K. 2004. Disruption of exonic splicing enhancer elements is the principal cause of exon skipping associated with seven nonsense or missense alleles of NF1. *Human Mutation,* 24**,** 491-501.

Zatz, M., De Paula, F., Starling, A. & Vainzof, M. 2003. The 10 autosomal recessive limb-girdle muscular dystrophies. *Neuromuscular disorders: NMD,* 13**,** 532-544.

Zellweger, H. N., E 1965. Central nervous system manifestastions in childhood muscular dystrophy (CMD). *Annales paediatrici. International review of pediatrics,* 205**,** 25 - 42.

Zhang, K., Qin, Z., Chen, T., Liu, J. S., Waterman, M. S. & Sun, F. 2005. HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics,* 21**,** 131-134.

6. Appendices

Appendix I: Homozygsity Mapping Results

Initial Dataset

Affected individuals. 5 minimum sequential homozygous SNPs

Length	Chrom	Marker (SNP name, position, nearest gene, genotypes for affecteds)
8	3	s04366.1(19896460)(no gene)[GG GG GG GG GG GG GG GG GG]
		s43586.1(19945849)(no gene)[GG GG GG GG GG GG GG GG GG]
		s44598.1(19989596)(no gene)[GG GG GG GG GG GG GG GG GG]
		s62248.1(20012335)(no gene)[AA AA AA AA AA AA AA AA AA]
		s39730.1(20027673)(Name=Uncharacterized)[GG GG GG GG GG GG GG GG GG]
		OAR3_21630699.1(20052991)(Name=PQLC3)[GG GG GG GG GG GG GG GG GG]
		OAR3_21684794.1(20105333)(Name=ROCK2)[AA AA AA AA AA AA AA AA AA]
		OAR3_21695741.1(20116222)(Name=ROCK2)[GG GG GG GG GG GG GG GG GG]
6	25	OAR25_19823535.1(19167503)(Name=NRBF2)[GG GG GG GG GG GG GG GG GG]
		OAR25_19926183.1(19270565)(Name=JMJD1C)[GG GG GG GG GG GG GG GG GG]
		OAR25_19941944.1(19286373)(Name=JMJD1C)[GG GG GG GG GG GG GG GG GG]
		OAR25_20026839.1(19369540)(Name=JMJD1C)[GG GG GG GG GG GG GG GG GG]
		OAR25_20041897.1(19386582)(Name=JMJD1C)[AA AA AA AA AA AA AA AA AA]
		OAR25_20106030.1(19438048)(Name=JMJD1C)[AA AA AA AA AA AA AA AA AA]
5	13	OAR13_60759835.1(55866953)(Name=CDH26)[AA AA AA AA AA AA AA AA AA]
		s47781.1(55915765)(no gene)[AA AA AA AA AA AA AA AA AA]
		OAR13_60821868.1(55934178)(no gene)[AA AA AA AA AA AA AA AA AA]
		OAR13_60855392.1(55970160)(Name=SYCP2)[AA AA AA AA AA AA AA AA AA]
		OAR13_60893851.1(56010691)(no gene)[AA AA AA AA AA AA AA AA AA]
5	11	s44351.1(49904840)(Name=CCDC57)[AA AA AA AA AA AA AA AA AA]
		s30936.1(49940471)(Name=FASN)[GG GG GG GG GG GG GG GG GG]
		s08804.1(49987461)(no gene)[GG GG GG GG GG GG GG GG GG]
		s49850.1(50004896)(no gene)[AA AA AA AA AA AA AA AA AA]

OAR11_53396507.1(50045164)(Name=LRRC45)[GG GG GG GG GG GG GG GG GG]

Carrier Individuals. 5 minimum sequential homozygous SNPs

9	18	OAR18_57753977.1(53977812)(Name=TGM5)[00 00 GG GG 00 00 00 00 GG GG GG 00 00]
		s59424.1(53988694)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		OAR18_57812263.1(54039497)(Name=TGM7)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		OAR18_57822087.1(54047823)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		s08497.1(54126281)(Name=TP53BP1)[00 00 GA GA 00 00 00 00 GA GA GA 00 00]
		s65227.1(54133053)(Name=TP53BP1)[00 00 GA GA 00 00 00 00 GA GA GA 00 00]
		OAR18_57945473.1(54174325)(Name=TP53BP1)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		s47566.1(54226355)(Name=MAP1B)[00 00 GG GG 00 00 00 00 GG GG GG 00 00]
		DU404011_458.1(54252674)(Name=PPIP5K1)[00 00 AG AG 00 00 00 00 AG AG AG 00 00]
8	8	OAR8_18721587.1(16842172)(no gene)[00 00 AG AG 00 00 00 00 AG AG AG 00 00]
		OAR8_18751992.1(16866718)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00]
		OAR8_18770427.1(16882313)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		OAR8_18777471.1(16888761)(Name=BROMI)[00 00 AG AG 00 00 00 00 AG AG AG 00 00]
		OAR8_18857269.1(16930039)(Name=BROMI)[00 00 CC CC 00 00 00 00 CC CC CC 00 00]
		OAR8_18904554.1(16975578)(Name=BROMI)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		OAR8_18948860.1(17014826)(Name=BROMI)[00 00 CA CA 00 00 00 00 CA CA CA 00 00]
		OAR8_18997620.1(17052479)(Name=BROMI)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
7	5	s02334.1(14617884)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		s57684.1(14686309)(Name=ANGPTL4)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		s17355.1(14712996)(Name=RAB11B)[00 00 AG AG 00 00 00 00 AG AG AG 00 00]
		OAR5_16996550.1(14775024)(Name=HNRNPM)[00 00 GA GA 00 00 00 00 GA GA GA 00 00]
		s15525.1(14792565)(Name=PRAM1)[00 00 AC AC 00 00 00 00 AC AC AC 00 00]
		OAR5_17075528.1(14847585)(Name=ADAMTS10)[00 00 AG AG 00 00 00 00 AG AG AG 00 00]
		OAR5_17102793.1(14879240)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00]
6	18	OAR18_33701857.1(32280143)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		OAR18_33752500.1(32330978)(Name=SIN3A)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
		OAR18_33774245.1(32352543)(Name=SIN3A)[00 00 GG GG 00 00 00 00 GG GG GG 00 00]
		OAR18_33835260.1(32409159)(Name=COMMD4)[00 00 GA GA 00 00 00 00 GA GA GA 00 00]
		s08407.1(32452463)(no gene)[00 00 AG AG 00 00 00 00 AG AG AG 00 00]

```
        OAR18_33902994.1(32479250)(no gene)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]

6   14  OAR14_3404872.1(3486435)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
        s23868.1(3609268)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
        OAR14_3688132.1(3712183)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
        s10865.1(3782575)(no gene)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]
        s57571.1(3800340)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
        s30966.1(3813758)(no gene)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]

6   9   OAR9_75040566.1(70840453)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
        OAR9_75075388.1(70876514)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
        OAR9_75093968.1(70895159)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
        s38039.1(70925830)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
        OAR9_75230004.1(70959064)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]
        OAR9_75289690.1(71053054)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]

6   1   OAR1_103393742.1(96909929)(Name=PDE4DIP)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
        OAR1_103420822.1(96934926)(Name=PDE4DIP)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]
        OAR1_103445194.1(96960192)(Name=PDE4DIP)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
        OAR1_103547224.1(97033682)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
        OAR1_103772253.1(97192711)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
        OAR1_103790218.1(97209577)(Name=FMO5)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]

5   23  OARX_125130516.1(100425156)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
        OARX_125084587.1(100474082)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
        OARX_125057668.1(100514641)(no gene)[ 00 00 00 AA 00 00 00 00 AA AA AA 00 00]
        OARX_125044988.1(100528609)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
        OARX_125031440.1(100546822)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]

5   18  OAR18_31865428.1(30609238)(no gene)[ 00 00 CA CA 00 00 00 00 CA CA CA 00 00]
        s04966.1(30658559)(Name=SCAPER)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
        OAR18_31946902.1(30682003)(Name=SCAPER)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
        OAR18_31974215.1(30706077)(Name=SCAPER)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
        s40729.1(30783420)(Name=SCAPER)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]

5   18  s24856.1(8815383)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
```

```
          OAR18_8693731.1(8913593)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR18_8733617.1(8953384)(Name=Uncharacterized)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]
          OAR18_8785465.1(9005514)(no gene)[ 00 00 00 AA 00 00 00 00 AA AA AA 00 00]
          OAR18_9078383.1(9031682)(no gene)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]

5     18  s40731.1(59640699)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s10572.1(59668774)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
          OAR18_63721952.1(59706427)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s36630.1(59743824)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s50855.1(59799782)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]

5     17  OAR17_28218413_X.1(25672666)(no gene)[ 00 00 GA 00 00 00 00 00 GA GA GA 00 00]
          OAR17_28261838.1(25713704)(no gene)[ 00 00 AC AC 00 00 00 00 AC AC AC 00 00]
          OAR17_28312619.1(25767191)(no gene)[ 00 00 AC AC 00 00 00 00 00 AC AC 00 00]
          OAR17_28363294.1(25826660)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
          OAR17_28390642.1(25862600)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]

5     16  OAR16_16777017.1(15134783)(Name=RNF180)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]
          OAR16_16801439.1(15159802)(Name=RNF180)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]
          OAR16_16902918.1(15280464)(Name=RNF180)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
          OAR16_16928494.1(15306397)(Name=RNF180)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR16_16984728.1(15362234)(no gene)[ 00 00 00 AA 00 00 00 00 AA AA AA 00 00]

5     16  OAR16_44680056.1(41134921)(Name=ZFR)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR16_44759148.1(41206143)(Name=MTMR12)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]
          OAR16_44812720.1(41229069)(Name=MTMR12)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          s58356.1(41258135)(Name=MTMR12)[ 00 00 AA 00 00 00 00 00 AA AA AA 00 00]
          OAR16_44884811.1(41295748)(no gene)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]

5     16  s45053.1(70559566)(Name=AHRR)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OARUn.54_510669.1(70664506)(Name=EXOC3)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          s53565.1(70710282)(Name=SLC9A3)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s51002.1(70900966)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          s37581.1(70922464)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]

5     14  s66108.1(34390195)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
```

```
            OAR14_35908377.1(34477621)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            s07117.1(34558103)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            s65863.1(34566935)(Name=GFOD2)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]
            OAR14_36030253_X.1(34596131)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]

5     13    s63569.1(21664236)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            OAR13_24163213.1(21673888)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            OAR13_24206262.1(21688556)(no gene)[ 00 00 CC 00 00 00 00 00 CC CC CC 00 00]
            s40227.1(21753962)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
            s07619.1(21806857)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
5     13    s37974.1(15256050)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
            s09489.1(15265028)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            OAR13_13397282.1(15307978)(Name=USP6NL)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            s23760.1(15359923)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            OAR13_13298202.1(15406880)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]

5     13    OAR13_60759835.1(55866953)(Name=CDH26)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            s47781.1(55915765)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            OAR13_60821868.1(55934178)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            OAR13_60855392.1(55970160)(Name=SYCP2)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            OAR13_60893851.1(56010691)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]

5     13    s67504.1(68158981)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
            s10589.1(68213972)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]
            OAR13_73412719.1(68241391)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]
            OAR13_73486401.1(68306113)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]
            OAR13_73527165.1(68347344)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]

5     11    s35197.1(30193254)(Name=DNAH9)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            s14669.1(30222632)(Name=ZNF18)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            OAR11_32129276.1(30250362)(Name=MAP2K4)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
            OAR11_32155360.1(30275940)(Name=MAP2K4)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
            OAR11_32204952.1(30329562)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]

5     11    OAR11_61713996.1(57193316)(Name=SLC39A11)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR11_61776871.1(57252925)(Name=Uncharacterized)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
```

```
          s07746.1(57283705)(Name=SLC39A11)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s65471.1(57292356)(Name=SLC39A11)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s01777.1(57367601)(Name=SLC39A11)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]

5     9   s36433.1(88690261)(Name=CNGB3)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR9_94270944_X.1(88745479)(Name=CPNE3)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR9_94310472.1(88775547)(Name=CPNE3)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s70574.1(88836294)(Name=WWP1)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR9_94375693.1(88900878)(Name=WWP1)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]

5     9   s28418.1(80781440)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR9_85509697.1(80838913)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR9_85560833.1(80885666)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]
          OAR9_85594743.1(80924591)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR9_85619253.1(80947471)(no gene)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]

5     6   OAR6_28048390.1(24659269)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR6_28075131.1(24687510)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR6_28132840.1(24714766)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR6_28175615.1(24755360)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR6_28396848.1(24968246)(Name=Uncharacterized)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]

5     4   OAR4_31672384.1(30120791)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR4_31690929.1(30147111)(no gene)[ 00 00 00 00 00 00 00 00 AG AG AG 00 00]
          OAR4_31727075.1(30179898)(no gene)[ 00 00 CA CA 00 00 00 00 CA CA CA 00 00]
          s01678.1(30220570)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]
          OAR4_31796734.1(30244105)(Name=SP4)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]

5     3   OAR3_6854899.1(6774118)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR3_6897820.1(6820713)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR3_6950846.1(6871259)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          s58010.1(6884511)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s13823.1(6989246)(Name=Uncharacterized)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]

5     3   OAR3_156362251.1(146504875)(Name=LRRK2)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR3_156409128.1(146552297)(Name=LRRK2)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
```

```
          OAR3_156474166.1(146621605)(Name=LRRK2)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR3_156484814_X.1(146631986)(Name=LRRK2)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
          OAR3_156555651.1(146703642)(Name=LRRK2)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]

5     3   s24023.1(60735953)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s51327.1(60804625)(Name=Uncharacterized)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]
          OAR3_64372580.1(60840244)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR3_64400364_X.1(60863772)(Name=UXS1)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR3_64436154.1(60899402)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]

5     2   s04563.1(248037266)(Name=RCC2)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          s43889.1(248063253)(Name=PADI6)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]
          s08754.1(248069420)(Name=PADI6)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          s45037.1(248103245)(Name=PADI3)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR2_262394560.1(248130533)(Name=PADI3)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]

5     2   OAR2_77928516.1(73142522)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR2_77962229.1(73188664)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR2_78002934_X.1(73226330)(Name=KIAA1432)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]
          OAR2_78055691.1(73287723)(Name=KIAA1432)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR2_78106813.1(73333474)(Name=KIAA1432)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]

5     2   s47673.1(88986269)(no gene)[ 00 00 CC CC 00 00 00 00 CC 00 CC 00 00]
          s40411.1(88992392)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00]
          OAR2_95405699.1(89076235)(Name=Uncharacterized)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR2_95050543.1(89189237)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]
          OAR2_95733865.1(89239384)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00]

5     2   s47616.1(83090209)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR2_88340779.1(83115753)(no gene)[ 00 00 CA CA 00 00 00 00 CA CA CA 00 00]
          s61655.1(83194854)(Name=TTC39B)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00]
          OAR2_88441803.1(83214641)(Name=TTC39B)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          s23687.1(83305173)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00]

5     1   s32216.1(128983286)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
          OAR1_139735768.1(129024743)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00]
```

OAR1_139783950.1(129072656)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00]
OAR1_140089069.1(129317012)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00]
OAR1_140104902.1(129332577)(no gene)[00 00 CA CA 00 00 00 00 CA CA CA 00 00]

Final Dataset

Affected Individuals. 5 minimum sequential homozygous SNPs

Length Chrom Marker (SNP name, position, nearest gene, genotypes for affecteds)

6 3 s44598.1(19989596)(no gene)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]
s62248.1(20012335)(no gene)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]
s39730.1(20027673)(Name=Uncharacterized)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]
OAR3_21630699.1(20052991)(Name=PQLC3)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]
OAR3_21684794.1(20105333)(Name=ROCK2)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]
OAR3_21695741.1(20116222)(Name=ROCK2)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]

5 25 OAR25_19926183.1(19270565)(Name=JMJD1C)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]
OAR25_19941944.1(19286373)(Name=JMJD1C)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]
OAR25_20026839.1(19369540)(Name=JMJD1C)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]
OAR25_20041897.1(19386582)(Name=JMJD1C)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]
OAR25_20106030.1(19438048)(Name=JMJD1C)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]

5 13 OAR13_60759835.1(55866953)(Name=CDH26)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]
s47781.1(55915765)(no gene)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]
OAR13_60821868.1(55934178)(no gene)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]
OAR13_60855392.1(55970160)(Name=SYCP2)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]
OAR13_60893851.1(56010691)(no gene)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]

5 11 s44351.1(49904840)(Name=CCDC57)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]
s30936.1(49940471)(Name=FASN)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]
s08804.1(49987461)(no gene)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]
s49850.1(50004896)(no gene)[AA AA AA AA AA AA AA AA AA AA AA AA AA AA]
OAR11_53396507.1(50045164)(Name=LRRC45)[GG GG GG GG GG GG GG GG GG GG GG GG GG GG]

Carrier individuals. 5 minimum sequential homozygous SNPs

6	14	OAR14_3404872.1(3486435)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		s23868.1(3609268)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		OAR14_3688132.1(3712183)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
		s10865.1(3782575)(no gene)[00 00 GA GA 00 00 00 00 GA GA GA 00 00 GA]
		s57571.1(3800340)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		s30966.1(3813758)(no gene)[00 00 GA GA 00 00 00 00 GA GA GA 00 00 GA]
6	9	OAR9_75040566.1(70840453)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		OAR9_75075388.1(70876514)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
		OAR9_75093968.1(70895159)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
		s38039.1(70925830)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		OAR9_75230004.1(70959064)(no gene)[00 00 CC CC 00 00 00 00 CC CC CC 00 00 CC]
		OAR9_75289690.1(71053054)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
5	23	OARX_125130516.1(100425156)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		OARX_125084587.1(100474082)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		OARX_125057668.1(100514641)(no gene)[00 00 00 AA 00 00 00 00 AA AA AA 00 00 AA]
		OARX_125044988.1(100528609)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
		OARX_125031440.1(100546822)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
5	18	s40731.1(59640699)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		s10572.1(59668774)(no gene)[00 00 AG AG 00 00 00 00 AG AG AG 00 00 AG]
		OAR18_63721952.1(59706427)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		s36630.1(59743824)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		s50855.1(59799782)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
5	16	s45053.1(70559566)(Name=AHRR)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		OARUn.54_510669.1(70664506)(Name=EXOC3)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
		s53565.1(70710282)(Name=SLC9A3)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
		s51002.1(70900966)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
		s37581.1(70922464)(no gene)[00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
5	14	s66108.1(34390195)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
		OAR14_35908377.1(34477621)(no gene)[00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]

```
          s07117.1(34558103)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          s65863.1(34566935)(Name=GFOD2)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00 CC]
          OAR14_36030253_X.1(34596131)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]

5     13  s67504.1(68158981)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00 AG]
          s10589.1(68213972)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00 CC]
          OAR13_73412719.1(68241391)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00 CC]
          OAR13_73486401.1(68306113)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00 CC]
          OAR13_73527165.1(68347344)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00 CC]

5     13  s37974.1(15256050)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          s09489.1(15265028)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR13_13397282.1(15307978)(Name=USP6NL)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          s23760.1(15359923)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR13_13298202.1(15406880)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]

5     13  s63569.1(21664236)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR13_24163213.1(21673888)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR13_24206262.1(21688556)(no gene)[ 00 00 CC 00 00 00 00 00 CC CC CC 00 00 CC]
          s40227.1(21753962)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00 AG]
          s07619.1(21806857)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]

5     13  OAR13_60759835.1(55866953)(Name=CDH26)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          s47781.1(55915765)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR13_60821868.1(55934178)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR13_60855392.1(55970160)(Name=SYCP2)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR13_60893851.1(56010691)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]

5     11  s35197.1(30193254)(Name=DNAH9)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          s14669.1(30222632)(Name=ZNF18)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR11_32129276.1(30250362)(Name=MAP2K4)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          OAR11_32155360.1(30275940)(Name=MAP2K4)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR11_32204952.1(30329562)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]

5     9   s28418.1(80781440)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          OAR9_85509697.1(80838913)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
```

```
          OAR9_85560833.1(80885666)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00 CC]
          OAR9_85594743.1(80924591)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR9_85619253.1(80947471)(no gene)[ 00 00 GA GA 00 00 00 00 GA GA GA 00 00 GA]

5     9   s36433.1(88690261)(Name=CNGB3)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR9_94270944_X.1(88745479)(Name=CPNE3)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR9_94310472.1(88775547)(Name=CPNE3)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          s70574.1(88836294)(Name=WWP1)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          OAR9_94375693.1(88900878)(Name=WWP1)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]

5     6   OAR6_28048390.1(24659269)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          OAR6_28075131.1(24687510)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR6_28132840.1(24714766)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          OAR6_28175615.1(24755360)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR6_28396848.1(24968246)(Name=Uncharacterized)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]

5     3   OAR3_6854899.1(6774118)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          OAR3_6897820.1(6820713)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          OAR3_6950846.1(6871259)(no gene)[ 00 00 AA AA 00 00 00 00 AA AA AA 00 00 AA]
          s58010.1(6884511)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          s13823.1(6989246)(Name=Uncharacterized)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]

5     2   s47673.1(88986269)(no gene)[ 00 00 CC CC 00 00 00 00 CC 00 CC 00 00 CC]
          s40411.1(88992392)(no gene)[ 00 00 AG AG 00 00 00 00 AG AG AG 00 00 AG]
          OAR2_95405699.1(89076235)(Name=Uncharacterized)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          OAR2_95050543.1(89189237)(no gene)[ 00 00 GG GG 00 00 00 00 GG GG GG 00 00 GG]
          OAR2_95733865.1(89239384)(no gene)[ 00 00 CC CC 00 00 00 00 CC CC CC 00 00 CC]
```

Appendix II: Top 100 SNPs by p-value in PLINK Initial Dataset

MARKER	CHR	BP	GENENAME	GENEBP	BPTOGENE	A1	A2	MODEL	FRQ_AFF	FRQ_UN	PVAL
s05374.1	16	22509031	(Close_to)_Name=MAP3K1	22392215..22419948	89083	A	G	REC	8/0	0/5	0.000312
OAR10_46125690.1	10	45580754	(Close_to)_Name=KLHL1	44869838..44933847	646907	A	G	REC	8/1	0/5	0.001281
OAR1_25842722.1	1	25561238	Name=EPS15	25557772..25839810		G	A	REC	8/1	0/5	0.001281
OAR12_83564723.1	12	75920876	Name=Uncharacterized	75883273..75923269		A	G	REC	8/1	0/5	0.001281
OAR16_29544100.1	16	27262705	(Close_to)_Name=ISL1	27720710..27730148	458005	G	A	REC	8/1	0/5	0.001281
OAR16_29864348.1	16	27555794	(Close_to)_Name=ISL1	27720710..27730148	164916	G	A	REC	8/1	0/5	0.001281
OAR23_60556779.1	23	56855903	(Close_to)_Name=ST8SIA3	56832406..56838701	17202	A	G	REC	8/1	0/5	0.001281
s67978.1	14	59364622	Name=SHISA7	59341240..59365691		A	G	REC	8/1	0/5	0.001281
s70688.1	23	21587590	(Close_to)_Name=Uncharacterized	21463393..21559101	28489	A	G	REC	8/1	0/5	0.001281
s71494.1	22	30158667	Name=RBM20	30135041..30358741		G	A	REC	8/1	0/5	0.001281
OAR10_42221999.1	10	41420084	(Close_to)_Name=PCDH9	40818672..40821770	598314	A	G	REC	9/0	1/4	0.001499
OAR2_195165011.1	2	184070090	Name=PTPN4	184051357..184394205		A	G	REC	9/0	1/4	0.001499
OAR3_222203288.1	3	206246224	(Close_to)_Name=FOXJ2	206246262..206314845	38	A	G	REC	9/0	1/4	0.001499
s64920.1	2	184492101	Name=RALB	184423271..184497145		G	A	REC	9/0	1/4	0.001499
s42749.1	5	106108323	Name=MAN2A1	105985035..106179347		G	A	REC	7/0	0/3	0.001565
OAR5_86798856.1	5	79001358	(Close_to)_Name=SSBP2	78722098..78807837	193521	C	A	REC	7/1	0/5	0.002078
DU240765_244.1	X	121439986	(Close_to)_Name=RIPPLY1	121429249..121431852	8134	A	G	REC	5/0	0/4	0.0027
DU498640_321.1	X	107865244	(Close_to)_Name=OCRL	107805408..107850626	14618	G	A	REC	5/0	0/4	0.0027
OARX_27076240.1	X	21770789	Name=PDK3	21723982..21775661		G	A	REC	5/0	0/4	0.0027
OAR1_268193141.1	1	248363482	Name=PIK3CB	248349664..248557797		A	G	REC	0/9	3/1	0.003054
OAR4_116232256.1	4	108742207	(Close_to)_Name=Uncharacterized	109039190..109039706	296983	A	G	REC	0/9	3/1	0.003054
s53402.1	1	263987408	Name=FTCD	263986324..264000997		A	G	REC	0/9	3/1	0.003054
s09120.1	17	54115859	Name=PPP1CC	54092664..54203976		A	G	REC	6/1	0/5	0.003415
OAR22_4602449.1	22	3698827	(Close_to)_Name=Uncharacterized	3560069..3560720	138107	G	A	REC	8/1	0/3	0.004678
OARX_29830880.1	X	22527863	(Close_to)_Name=PPP2R1A	22623303..22625072	95440	G	A	REC	5/0	0/3	0.004678
OAR10_16100262.1	10	17223259	(Close_to)_Name=HTR2A	17135647..17196555	26704	G	A	REC	7/2	0/5	0.005289
OAR1_135764540.1	1	125198643	Name=TAK1L	125169912..125261167		C	A	REC	7/2	0/5	0.005289
OAR1_38581202.1	1	37533058	Name=DOCK7	37492256..37712875		G	A	REC	7/2	0/5	0.005289
OAR21_11273092_X.1	21	9756580	(Close_to)_Name=Uncharacterized	9853739..9854363	97159	C	A	REC	7/2	0/5	0.005289
OAR21_16900072.1	21	14850333	(Close_to)_Name=Uncharacterized	15153720..15223232	303387	G	A	REC	7/2	0/5	0.005289
OAR21_50593433.1	21	45569222	Name=IGHMBP2	45566574..45572457		A	C	REC	7/2	0/5	0.005289

OAR21_50593433.1	21	45569222	Name=MRPL21	45537653..45594155		A	C	REC	7/2	0/5	0.005289
OAR2_219701580.1	2	207476786	Name=ADAM23	207429120..207617493		G	A	REC	7/2	0/5	0.005289
OAR5_14147739.1	5	11978651	(Close_to)_Name=Uncharacterized	11962085..11962632	16019	G	A	REC	7/2	0/5	0.005289
OAR7_95578007.1	7	87876080	Name=NRXN3	87678336..88105158		G	A	REC	7/2	0/5	0.005289
s27070.1	10	34319770	(Close_to)_Name=NUPL1	34091325..34247831	71939	A	G	REC	7/2	0/5	0.005289
s36676.1	14	58909056	Name=CACNG7	58865906..58932185		A	G	REC	7/2	0/5	0.005289
s56510.1	22	16155452	Name=SORBS1	16075857..16209406		A	G	REC	7/2	0/5	0.005289
s73257.1	14	59440211	Name=BRSK1	59433357..59452837		G	A	REC	7/2	0/5	0.005289
s73257.1	14	59440211	Name=HSPBP1	59422831..59475484		G	A	REC	7/2	0/5	0.005289
OAR2_162921437.1	2	153751525	(Close_to)_Name=KCNJ3	153620397..153716544	34981	G	A	REC	7/2	0/3	0.007215
OAR25_43581009.1	25	41260077	(Close_to)_Name=GLUD1	41212024..41249846	10231	G	A	REC	7/0	1/3	0.007215
OAR3_51396190.1	3	48234418	(Close_to)_Name=CFDP2	48260114..48316549	25696	G	A	REC	7/1	0/3	0.007215
s74404.1	20	49767664	(Close_to)_Name=Uncharacterized	49761673..49762401	5263	T	A	REC	7/1	0/3	0.007215
s74865.1	22	31081205	(Close_to)_Name=Uncharacterized	31405872..31422997	324667	G	A	REC	5/0	0/2	0.008151
OAR10_40339520.1	10	39489315	(Close_to)_Name=PCDH9	39646385..39695585	157070	G	A	REC	9/0	2/3	0.008752
OAR10_49726084.1	10	48932289	(Close_to)_Name=KLF12	48967445..49325568	35156	A	G	REC	9/0	2/3	0.008752
OAR10_49941391.1	10	49105942	Name=KLF12	48967445..49325568		C	A	REC	9/0	2/3	0.008752
OAR10_89499060.1	10	81950954	(Close_to)_Name=Uncharacterized	81839197..81854151	96803	A	C	REC	0/9	3/2	0.008752
OAR11_57971056.1	11	54154637	Name=METTL23	54049613..54217476		A	G	REC	0/9	3/2	0.008752
OAR11_63539095.1	11	58815143	(Close_to)_Name=Uncharacterized	59402248..59402739	587105	G	A	REC	0/9	3/2	0.008752
OAR12_56357383.1	12	50938436	(Close_to)_Name=Uncharacterized	50852326..50853058	85378	G	A	REC	9/0	2/3	0.008752
OAR12_65173671.1	12	58752009	(Close_to)_Name=SOAT1	58686926..58751930	79	G	A	REC	0/9	3/2	0.008752
OAR1_268175642.1	1	248348219	(Close_to)_Name=PIK3CB	248349664..248557797	1445	A	G	REC	0/9	3/2	0.008752
OAR1_275446990.1	1	254866909	(Close_to)_Name=NPHP3	254936877..254990938	69968	A	G	REC	0/9	3/2	0.008752
OAR12_82572003.1	12	74953097	Name=C1ORF53	74952060..74953570		A	G	REC	0/9	3/2	0.008752
OAR1_287160585.1	1	265450251	(Close_to)_Name=Uncharacterized	265771034..265771240	320783	C	A	REC	0/9	3/2	0.008752
OAR14_17682700.1	14	17098867	(Close_to)_Name=Uncharacterized	16993407..17093126	5741	G	A	REC	0/9	3/2	0.008752
OAR14_29971688.1	14	28809538	Name=CDH8	28745512..29140552		A	G	REC	0/9	3/2	0.008752
OAR15_61453696.1	15	56087018	(Close_to)_Name=RPS3A	56176897..56177691	89879	A	G	REC	0/9	3/2	0.008752
OAR1_65172267.1	1	61589657	Name=SYDE2	61571675..61610559		C	A	REC	0/9	3/2	0.008752
OAR17_45018430.1	17	41713625	Name=GRIA2	41518585..41716171		A	G	REC	0/9	3/2	0.008752
OAR1_77436037.1	1	72378271	(Close_to)_Name=PTBP2	73057054..73115320	678783	A	G	REC	0/9	3/2	0.008752
OAR21_13977432.1	21	12314142	(Close_to)_Name=PRCP	12226040..12313295	847	G	A	REC	0/9	3/2	0.008752
OAR21_1404105.1	21	1081098	(Close_to)_Name=CCDC67	899477..1027156	53942	G	A	REC	0/9	3/2	0.008752

OAR21_17340576.1	21	15255502	(Close_to)_Name=Uncharacterized	15153720..15223232	32270	G	A	REC	0/9	3/2	0.008752
OAR21_36349971.1	21	32729749	(Close_to)_Name=Uncharacterized	32905142..32906130	175393	A	G	REC	9/0	2/3	0.008752
OAR21_8713943.1	21	7375018	(Close_to)_Name=BNIP3L	7504274..7504831	129256	A	G	REC	0/9	3/2	0.008752
OAR23_27969114_X.1	23	26831850	(Close_to)_Name=Uncharacterized	27076289..27076894	244439	G	A	REC	0/9	3/2	0.008752
OAR23_36959489.1	23	34992397	Name=GREB1L	34979403..34995122		G	A	REC	0/9	3/2	0.008752
OAR23_58632679.1	23	55141170	Name=TCF4	54803980..55179189		C	A	REC	0/9	3/2	0.008752
OAR23_59623047.1	23	55980030	Name=Uncharacterized	55956401..56028668		A	G	REC	0/9	3/2	0.008752
OAR3_104479755.1	3	98216780	(Close_to)_Name=Uncharacterized	98044428..98123319	93461	G	A	REC	0/9	3/2	0.008752
OAR3_119856323.1	3	112419045	(Close_to)_Name=ZDHHC17	112510300..112590526	91255	G	A	REC	0/9	3/2	0.008752
OAR3_123897974.1	3	116215835	Name=PTPRQ	116104058..116578874		G	A	REC	0/9	3/2	0.008752
OAR3_124547379.1	3	116846794	(Close_to)_Name=Uncharacterized	116797892..116841830	4964	C	A	REC	0/9	3/2	0.008752
OAR3_85112203.1	3	80551030	(Close_to)_Name=PLEKHH2	80437970..80536838	14192	A	G	REC	0/9	3/2	0.008752
OAR3_89528817.1	3	84570615	(Close_to)_Name=Uncharacterized	84673885..84681092	103270	G	A	REC	0/9	3/2	0.008752
OAR3_99622529.1	3	93811641	(Close_to)_Name=Uncharacterized	94005707..94008560	194066	G	A	REC	0/9	3/2	0.008752
OAR4_115857083.1	4	108409074	(Close_to)_Name=Uncharacterized	109039190..109039706	630116	A	C	REC	0/9	3/2	0.008752
OAR4_116002501.1	4	108627193	(Close_to)_Name=Uncharacterized	109039190..109039706	411997	G	A	REC	0/9	3/2	0.008752
OAR4_116082708.1	4	108574116	(Close_to)_Name=Uncharacterized	109039190..109039706	465074	G	A	REC	0/9	3/2	0.008752
OAR6_124646989_X.1	6	109697825	(Close_to)_Name=CPEB2	109725908..109793545	28083	A	G	REC	9/0	2/3	0.008752
OAR6_124660937.1	6	109711650	(Close_to)_Name=CPEB2	109725908..109793545	14258	G	A	REC	9/0	2/3	0.008752
OAR7_12244180.1	7	11962077	(Close_to)_Name=PARP16	11981614..12005480	19537	A	G	REC	0/9	3/2	0.008752
OAR8_40528238.1	8	37619396	(Close_to)_Name=C6ORF168	37512147..37585064	34332	A	G	REC	0/9	3/2	0.008752
OAR8_48775633.1	8	45269769	(Close_to)_Name=Uncharacterized	44284182..44306552	963217	A	G	REC	0/9	3/2	0.008752
OAR8_50190157.1	8	46706605	(Close_to)_Name=MAP3K7	46809660..46880280	103055	C	A	REC	9/0	2/3	0.008752
OAR8_56910466.1	8	53042000	Name=THEMIS	52929118..53133576		A	G	REC	0/9	3/2	0.008752
OAR8_58134022.1	8	54236077	Name=LAMA2	54216178..54895696		A	G	REC	0/9	3/2	0.008752
OAR8_58181648_X.1	8	54280168	Name=LAMA2	54216178..54895696		A	G	REC	0/9	3/2	0.008752
OAR8_75882266.1	8	70709943	(Close_to)_Name=GRM1	70557520..70670365	39578	G	A	REC	0/9	3/2	0.008752
OAR8_84067219.1	8	77949827	(Close_to)_Name=RPS6	77948039..77948788	1039	A	G	REC	0/9	3/2	0.008752
OAR8_91938724.1	8	85260120	(Close_to)_Name=PACRG	85277553..85380230	17433	C	A	REC	0/9	3/2	0.008752
s03652.1	17	41855593	Name=GLRB	41765919..41856956		A	G	REC	0/9	3/2	0.008752
s04619.1	23	59008271	(Close_to)_Name=Uncharacterized	59023858..59024268	15587	G	A	REC	0/9	3/2	0.008752
s19512.1	3	76037302	(Close_to)_Name=FOXN2	76004936..76026322	10980	G	A	REC	0/9	3/2	0.008752
s20069.1	8	79423092	(Close_to)_Name=HADH	79574321..79574761	151229	G	A	REC	0/9	3/2	0.008752
s23660.1	8	63246801	Name=NHSL1	63234873..63382699		A	G	REC	0/9	3/2	0.008752

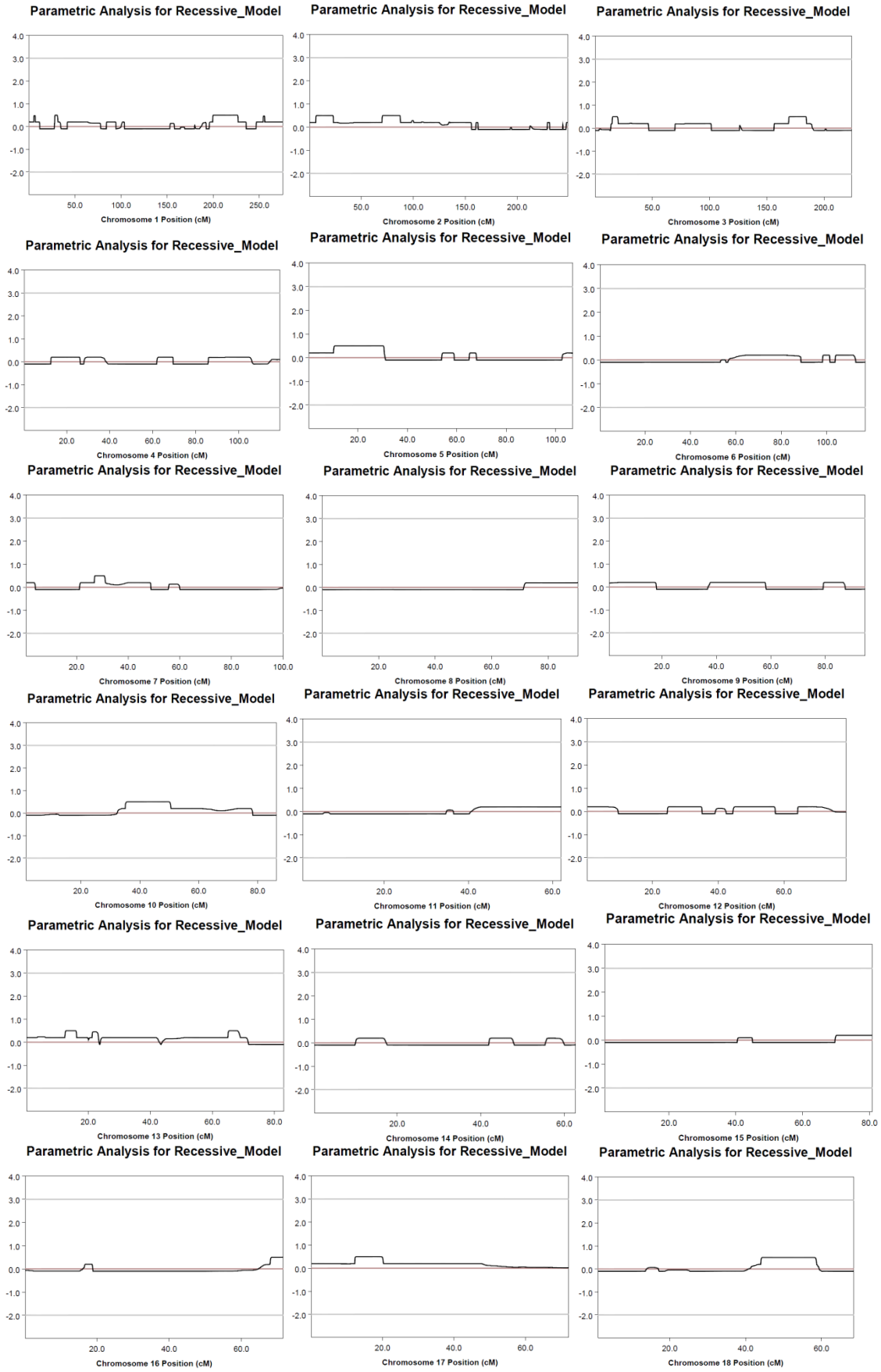
Appendix III: Top 100 SNPs by p-value in PLINK Final Dataset

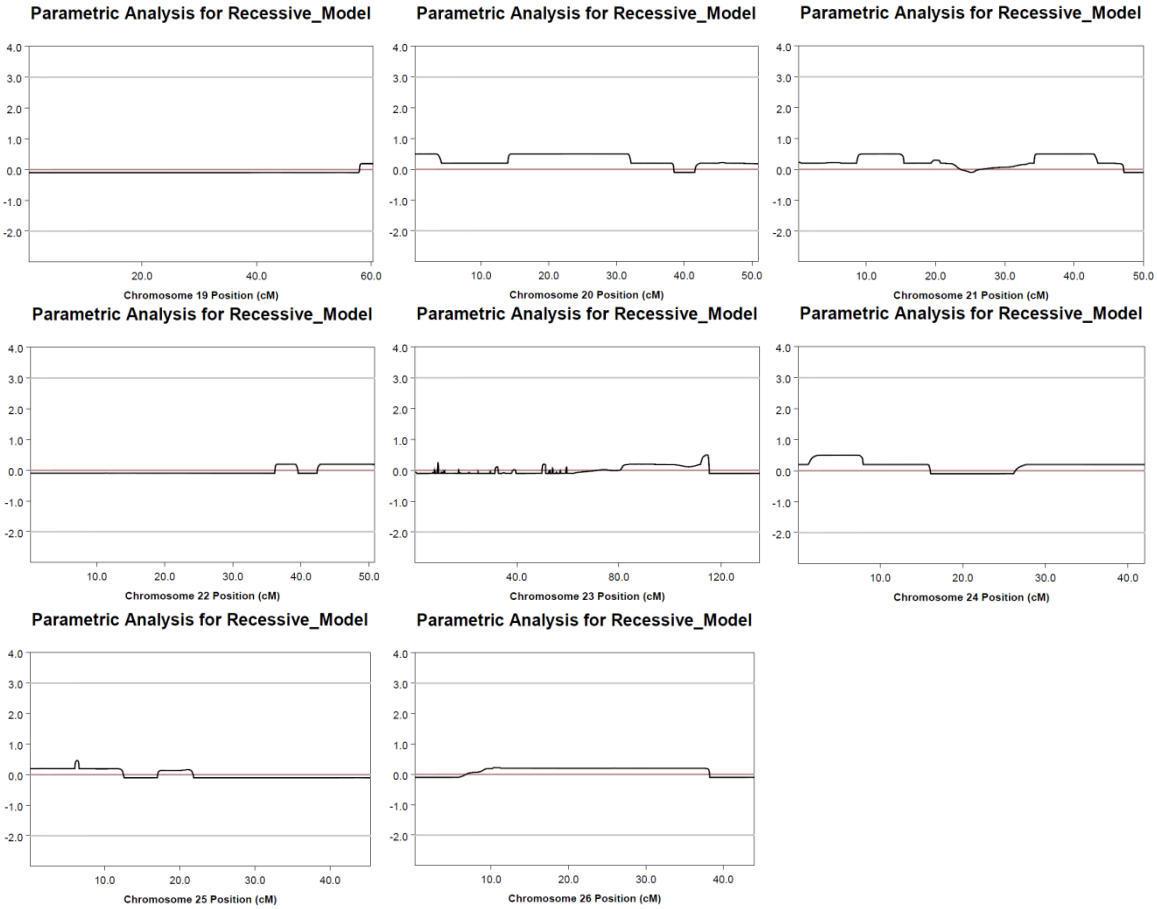
MARKER	CHR	BP	GENENAME	GENEBP	BPTOGENE	A1	A2	MODEL	FRQ_AFF	FRQ_UN	PVAL
OAR12_83564723.1	12	75920876	Name=Uncharacterized	75883273..75923269		A	G	REC	12/2	0/6	0.000336
OAR23_60556779.1	23	56855903	(Close_to)_Name=ST8SIA3	56832406..56838701	17202	A	G	REC	12/2	0/6	0.000336
OAR5_14147739.1	5	11978651	(Close_to)_Name=Uncharacterized	11962085..11962632	16019	G	A	REC	12/2	0/6	0.000336
OAR25_43581009.1	25	41260077	(Close_to)_Name=GLUD1	41212024..41249846	10231	G	A	REC	12/0	41365	0.000395
s05374.1	16	22509031	(Close_to)_Name=MAP3K1	22392215..22419948	89083	A	G	REC	11/2	0/6	0.000516
OAR10_49726084.1	10	48932289	(Close_to)_Name=KLF12	48967445..49325568	35156	A	G	REC	14/0	2/4	0.000636
OAR21_36349971.1	21	32729749	(Close_to)_Name=Uncharacterized	32905142..32906130	175393	A	G	REC	14/0	2/4	0.000636
OAR2_195165011.1	2	1.84E+08	Name=PTPN4	184051357..184394205		A	G	REC	14/0	2/4	0.000636
OAR6_124646989_X.1	6	1.1E+08	(Close_to)_Name=CPEB2	109725908..109793545	28083	A	G	REC	14/0	2/4	0.000636
OAR6_124660937.1	6	1.1E+08	(Close_to)_Name=CPEB2	109725908..109793545	14258	G	A	REC	14/0	2/4	0.000636
DU498640_321.1	23	1.08E+08	(Close_to)_Name=SERPINB10	62299681..62326225	45539019	G	A	REC	13/1	1/5	0.000656
OAR10_42221999.1	10	41420084	(Close_to)_Name=PCDH9	40818672..40821770	598314	A	G	REC	13/1	1/5	0.000656
OARX_27076240.1	23	21770789	(Close_to)_Name=GALNT1	21679987..21748387	22402	G	A	REC	13/1	1/5	0.000656
OAR10_16100262.1	10	17223259	(Close_to)_Name=HTR2A	17135647..17196555	26704	G	A	REC	11/3	0/6	0.001209
OAR21_11273092_X.1	21	9756580	(Close_to)_Name=Uncharacterized	9853739..9854363	97159	C	A	REC	11/3	0/6	0.001209
OAR21_16900072.1	21	14850333	(Close_to)_Name=Uncharacterized	15153720..15223232	303387	G	A	REC	11/3	0/6	0.001209
OAR24_37064565.1	24	33961710	(Close_to)_Name=CCL26	33952957..33957507	4203	G	A	REC	11/3	0/6	0.001209
OARX_29188769.1	23	19623665	Name=Uncharacterized	19583942..19640640		G	A	REC	11/3	0/6	0.001209
s27070.1	10	34319770	(Close_to)_Name=NUPL1	34091325..34247831	71939	A	G	REC	11/3	0/6	0.001209
s43015.1	24	33953578	Name=CCL26	33952957..33957507		G	A	REC	11/3	0/6	0.001209
OAR3_193744872.1	3	1.8E+08	Name=IFT27	179967963..179990644		G	A	REC	10/3	0/6	0.001799
s10237.1	24	24989534	Name=IL21R	24987787..25002160		A	C	REC	10/3	0/6	0.001799
s09120.1	17	54115859	Name=PPP1CC	54092664..54203976		A	G	REC	10/1	1/5	0.002205
OAR26_14049024.1	26	11375662	(Close_to)_Name=ODZ3	12081624..12141082	705962	C	A	REC	11/3	0/5	0.002254
OAR5_10935525.1	5	9535742	(Close_to)_Name=CD97	9450503..9467533	68209	A	G	REC	11/3	0/5	0.002254
OAR10_25624621.1	10	25709520	(Close_to)_Name=Uncharacterized	25678169..25678828	30692	A	G	REC	12/2	1/5	0.00301
OAR10_33307197.1	10	33021724	Name=Uncharacterized	32941724..33027426		G	A	REC	12/2	1/5	0.00301
OAR10_33338187.1	10	33053891	(Close_to)_Name=Uncharacterized	32941724..33027426	26465	G	A	REC	12/2	1/5	0.00301
OAR10_58058749.1	10	56914955	(Close_to)_Name=NFYB	56251109..56270332	644623	A	G	REC	12/2	1/5	0.00301
OAR10_69032148.1	10	66779341	Name=GPC5	66542081..66793149		G	A	REC	12/2	1/5	0.00301
OAR1_25842722.1	1	25561238	Name=EPS15	25557772..25839810		G	A	REC	12/2	1/5	0.00301

s06239.1	24	9352754	(Close_to)_Name=TEKT5	9359214..9407395	6460	A	G	REC	12/2	1/5	0.00301
s43183.1	10	51411639	(Close_to)_Name=LMO7	51117948..51358374	53265	A	G	REC	12/2	1/5	0.00301
s51842.1	25	41186480	(Close_to)_Name=SNCG	41170873..41183542	2938	G	A	REC	12/2	1/5	0.00301
s67978.1	14	59364622	Name=SHISA7	59341240..59365691		A	G	REC	12/2	1/5	0.00301
OAR10_29907137.1	10	29882404	(Close_to)_Name=B3GALTL	29893792..29994174	11388	A	G	REC	11/1	1/4	0.003128
OAR10_25988324.1	10	26056792	Name=NBEA	26011868..26592547		A	G	REC	10/4	0/6	0.003415
OAR12_10816504.1	12	8708988	(Close_to)_Name=Uncharacterized	8511596..8512941	196047	A	G	REC	10/4	0/6	0.003415
OAR16_31129400.1	16	28754634	(Close_to)_Name=EMB	28692740..28753002	1632	A	G	REC	10/4	0/6	0.003415
OAR21_10442989.1	21	9036537	(Close_to)_Name=PICALM	8950524..9036200	337	A	G	REC	10/4	0/6	0.003415
OAR21_10461107.1	21	9054725	(Close_to)_Name=PICALM	8950524..9036200	18525	C	A	REC	10/4	0/6	0.003415
OAR24_30139569_X.1	24	27613295	Name=MRPS17	27612710..27636615		G	A	REC	10/4	0/6	0.003415
OAR24_30146533.1	24	27620251	Name=MRPS17	27612710..27636615		A	G	REC	10/4	0/6	0.003415
OAR3_182795446.1	3	1.7E+08	(Close_to)_Name=UTP20	170228696..170315908	29193	A	C	REC	10/4	0/6	0.003415
OAR3_19616550_X.1	3	18106309	Name=KIDINS220	18105955..18190292		A	G	REC	10/4	0/6	0.003415
OAR3_25097037.1	3	23336791	(Close_to)_Name=NBAS	23499495..23835462	162704	C	A	REC	10/4	0/6	0.003415
s14379.1	3	1.8E+08	(Close_to)_Name=IFT27	179967963..179990644	1554	G	A	REC	10/4	0/6	0.003415
s27590.1	1	2.64E+08	(Close_to)_Name=S100B	264414341..264417467	9247	G	A	REC	10/4	0/6	0.003415
s56268.1	10	25943893	(Close_to)_Name=Uncharacterized	25954017..25970895	10124	G	A	REC	10/4	0/6	0.003415
s60222.1	23	45564934	(Close_to)_Name=SLC14A2	45567420..45633026	2486	G	A	REC	10/4	0/6	0.003415
s60732.1	1	2.68E+08	(Close_to)_Name=Uncharacterized	268094342..268197646	38051	G	A	REC	10/4	0/6	0.003415
s61799.1	10	30924195	(Close_to)_Name=UBL3	31052832..31095668	128637	G	A	REC	10/4	0/6	0.003415
s70688.1	23	21587590	(Close_to)_Name=Uncharacterized	21463393..21559101	28489	A	G	REC	10/4	0/6	0.003415
s71494.1	22	30158667	Name=RBM20	30135041..30358741		G	A	REC	10/4	0/6	0.003415
OAR14_68222343.1	14	61709350	(Close_to)_Name=Uncharacterized	61693493..61697083	12267	A	C	REC	8/3	0/6	0.004092
OAR10_5693105.1	10	7492074	(Close_to)_Name=PCDH17	5438145..5545654	1946420	C	A	REC	0/14	3/3	0.004108
OAR10_87392185.1	10	80078689	(Close_to)_Name=Uncharacterized	79988268..79991729	86960	A	C	REC	0/14	3/3	0.004108
OAR15_61453696.1	15	56087018	(Close_to)_Name=RPS3A	56176897..56177691	89879	A	G	REC	0/14	3/3	0.004108
OAR15_72854094.1	15	67326393	(Close_to)_Name=Uncharacterized	68406735..68410135	1080342	A	G	REC	0/14	3/3	0.004108
OAR16_29788902.1	16	27479191	(Close_to)_Name=ISL1	27720710..27730148	241519	G	A	REC	0/14	3/3	0.004108
OAR16_32987078.1	16	30357140	(Close_to)_Name=FGF10	30468800..30562188	111660	G	A	REC	0/14	3/3	0.004108
OAR16_63260453_X.1	16	57955217	(Close_to)_Name=ANKH	58190075..58357213	234858	G	A	REC	0/14	3/3	0.004108
OAR16_71436671.1	16	65681259	Name=ADCY2	65280195..65721159		A	C	REC	0/14	3/3	0.004108
OAR17_45018430.1	17	41713625	Name=GRIA2	41518585..41716171		A	G	REC	0/14	3/3	0.004108
OAR17_75701731.1	17	69462601	Name=INPP5J	69374736..69488557		A	C	REC	0/14	3/3	0.004108

OAR18_16433946.1	18	16365432	Name=AGBL1	16105736..16817080		G	A	REC	0/14	3/3	0.004108
OAR23_36959489.1	23	34992397	Name=GREB1L	34979403..34995122		G	A	REC	0/14	3/3	0.004108
OAR3_35192406.1	3	32861548	Name=DTNB	32602005..32903617		G	A	REC	0/14	3/3	0.004108
OAR4_26645098.1	4	25358759	Name=TSPAN13	25327414..25364818		A	G	REC	0/14	3/3	0.004108
OAR6_123771724.1	6	1.09E+08	(Close_to)_Name=Uncharacterized	109032515..109032766	160073	A	G	REC	0/14	3/3	0.004108
OAR6_126093150.1	6	1.11E+08	(Close_to)_Name=Uncharacterized	110908934..110910259	119185	G	A	REC	0/14	3/3	0.004108
OAR8_40528238.1	8	37619396	(Close_to)_Name=C6ORF168	37512147..37585064	34332	A	G	REC	0/14	3/3	0.004108
OAR8_47582140.1	8	44115521	(Close_to)_Name=EPHA7	44013118..44048755	66766	A	C	REC	0/14	3/3	0.004108
OAR8_48775633.1	8	45269769	(Close_to)_Name=Uncharacterized	44284182..44306552	963217	A	G	REC	0/14	3/3	0.004108
OAR8_50190157.1	8	46706605	(Close_to)_Name=MAP3K7	46809660..46880280	103055	C	A	REC	14/0	3/3	0.004108
OAR8_84067219.1	8	77949827	(Close_to)_Name=RPS6	77948039..77948788	1039	A	G	REC	0/14	3/3	0.004108
s03652.1	17	41855593	Name=GLRB	41765919..41856956		A	G	REC	0/14	3/3	0.004108
s10803.1	6	1.1E+08	(Close_to)_Name=Uncharacterized	109516618..109616046	10377	A	G	REC	0/14	3/3	0.004108
s57209.1	6	1.11E+08	Name=LDB2	111166384..111261472		C	A	REC	0/14	3/3	0.004108
s65924.1	16	17453973	(Close_to)_Name=Uncharacterized	17095578..17096622	357351	A	G	REC	0/14	3/3	0.004108
s67019.1	9	94116528	Name=COL14A1	93985570..94217441		A	G	REC	0/14	3/3	0.004108
s60527.1	3	30648203	(Close_to)_Name=Uncharacterized	30552647..30582515	65688	A	C	REC	11/2	1/5	0.004316
s67652.1	24	3439178	Name=SRL	3423050..3457451		A	G	REC	11/2	1/5	0.004316
OAR10_40339520.1	10	39489315	(Close_to)_Name=PCDH9	39646385..39695585	157070	G	A	REC	13/1	2/4	0.004845
OAR10_49941391.1	10	49105942	Name=KLF12	48967445..49325568		C	A	REC	13/1	2/4	0.004845
OAR10_50744611.1	10	49914859	(Close_to)_Name=N4BP2	49655604..49656128	258731	G	A	REC	13/1	2/4	0.004845
OAR10_60925663.1	10	59463069	(Close_to)_Name=SLITRK1	59237084..59242396	220673	A	C	REC	13/1	2/4	0.004845
OAR10_71498684.1	10	69161897	(Close_to)_Name=GPC6	69280455..69417471	118558	G	A	REC	13/1	2/4	0.004845
OAR11_14058741.1	11	14184155	Name=TAF15	14145853..14215694		A	G	REC	13/1	2/4	0.004845
OAR11_14058741.1	11	14184155	Name=MMP28	14174664..14212927		A	G	REC	13/1	2/4	0.004845
OAR1_265795032.1	1	2.46E+08	(Close_to)_Name=RPS19	246180197..246180640	43573	G	A	REC	13/1	2/4	0.004845
OAR1_77436037.1	1	72378271	(Close_to)_Name=PTBP2	73057054..73115320	678783	A	G	REC	1/13	4/2	0.004845
OAR2_194104000.1	2	1.83E+08	(Close_to)_Name=Uncharacterized	183101112..183110042	40730	C	A	REC	13/1	2/4	0.004845
OAR23_59623047.1	23	55980030	Name=Uncharacterized	55956401..56028668		A	G	REC	1/13	4/2	0.004845
OAR3_104479755.1	3	98216780	(Close_to)_Name=Uncharacterized	98044428..98123319	93461	G	A	REC	1/13	4/2	0.004845
OAR3_119856323.1	3	1.12E+08	(Close_to)_Name=ZDHHC17	112510300..112590526	91255	G	A	REC	1/13	4/2	0.004845
OAR3_123897974.1	3	1.16E+08	Name=PTPRQ	116104058..116578874		G	A	REC	1/13	4/2	0.004845
OAR3_124547379.1	3	1.17E+08	(Close_to)_Name=Uncharacterized	116797892..116841830	4964	C	A	REC	1/13	4/2	0.004845
OAR3_99622529.1	3	93811641	(Close_to)_Name=Uncharacterized	94005707..94008560	194066	G	A	REC	1/13	4/2	0.004845

Appendix IV: LOD Score Distribution by Chromosome Using Merlin




Parametric Analysis for Recessive_Model
Chromosome 19 Position (cM)
Parametric Analysis for Recessive_Model
Chromosome 20 Position (cM)
Parametric Analysis for Recessive_Model
Chromosome 21 Position (cM)
Parametric Analysis for Recessive_Model
Chromosome 22 Position (cM)
Parametric Analysis for Recessive_Model
Chromosome 23 Position (cM)
Parametric Analysis for Recessive_Model
Chromosome 24 Position (cM)
Parametric Analysis for Recessive_Model
Chromosome 25 Position (cM)
Parametric Analysis for Recessive_Model
Chromosome 26 Position (cM)

Appendix V: Perl source code for Homozygosity_Mapper

Please note that font size has been decreased to assist with layout readability of the code

```
#!/usr/bin/perl

# Authors: Jez Supreme and Kim Carter, 2013
# Purpose: Script to identify runs of homozygous SNP markers from linkage formatted data
#
# Inputs:  1. .ped text fle, in standard linkage ped format
#            2. .map text file, in standard linkage map format
#          3. .dat text file, in standard linkage dat format
#          4. affection status to filter in ped file ie 2=only affecteds, 1=only carriers/controls
#            5. number of mismatches (genotypes) allowed within a run of HZ
#            6. cutoff filter to reduce outputs ie minimum length of HZ runs of SNPs to display

# Outputs:  Text output of HZ regions, including list of markers and genomic locations of the composite SNPs


use strict;

#locate of sheep genome annotation file
my $gff = "/data/gff_files/Oarv3.1.protein.gene.gff3";

#check correct arguments are supplied
if ($#ARGV!=5) {
          print "Usage: Usage: get_homozyg.pl <input file.ped> <input file.map> <input file.dat> <aff status> <number mismatch per SNP> <minimum cutoff HZ run>\n";
          exit;
}

#declare variables
my $mapfilecol = 3; # ie 4 colum map
my $pedfile = $ARGV[0];
my $mapfile = $ARGV[1];
my $datfile = $ARGV[2];
my $aff = $ARGV[3];
my $mismatch = $ARGV[4];
my $cutoff = $ARGV[5];
my %results = ();
my %start = ();
my %end = ();
my %names = ();

#read GFF file into memory
open(GFF, "<$gff") || die("Failed to open $gff for reading");
while(<GFF>)
{
          my $line = $_;
          chomp($line);

          $line =~ s/\n//g;  # \n
          $line =~ s/\r//g;  # \r

          my @cols = split(/\t/,$line); # split it

          if ($cols[2] eq "mRNA") # only save features if they are MRNA ie gene features
          {
                    #scaffold004293     GLEAN     mRNA      16        1497      0.999942  -         .
          D=CCG000001.0;Name=Uncharacterized;Homology=B4AF47_BACPU;Note=BLAST.hit.in.refproteindatabaseB4AF47

                    #extract name and chromsome from annotation
                    my $namebit = $cols[8];
                    my @ns = split(/;/,$namebit); # split it

                    my $chr = $cols[0];
                    #OAR11
```

```
                    if ($chr =~ m/OAR/)
                    {
                              $chr = substr($chr,3);
                    }


                    #save gene start and end positions
                    if (exists($start{$chr}{$cols[3]}))
                    {
                              $start{$chr}{$cols[3]} = $start{$chr}{$cols[3]}."::".$ns[1];

                    }
                    else
                    {
                              $start{$chr}{$cols[3]} = $ns[0];

                    }

                    $end{$chr}{$ns[0]} = $cols[4];

                    #print "Saving: $chr, $cols[3]..$cols[4] as $ns[0]\n";
                    $names{$chr}{$cols[3]}{$cols[4]} = $ns[1];
          }
}
close(GFF);


#read DAT file into memory
my %datmap = ();
my $mcount=-1;
open(COLS, "<$datfile") || die("Failed to open $datfile for reading");
while(<COLS>)
{

          my $line = $_;
          chomp($line);

          #trim end of line characters
          $line =~ s/\n//g;  # \n
          $line =~ s/\r//g;  # \r
        $line =~ s/\t/ /g;
          $line =~ s/^\s+//;
          $line =~ s/\s+$//;

          if ($line =~/^M /)
          {
                    $mcount++;

                    #save marker
                    # (6 + (mnumber*2)) and (6 + (mcount*2)+1)
                    my @cols = split(/ /,$line); # split it

                    # datmap{colun number in ped file} -> snp name
                    $datmap{(6+($mcount*2))} = $cols[1];  #get name for column
                    #print "storing ".(6+($mcount*2))." as $cols[1] and $cols[0]\n";
          }

}
close(COLS);

#read map file into memory
my %markerinfo = ();
open(COLS, "<$mapfile") || die("Failed to open $mapfile for reading");
while(<COLS>)
{
          my $line = $_;
          chomp($line);

          #trim end of line characters
          $line =~ s/\n//g;  # \n
```

```
          $line =~ s/\r//g;  # \r
        $line =~ s/\t/ /g;
          $line =~ s/^\s+//;
          $line =~ s/\s+$//;

          #save marker location - chromosome and position
          my @cols = split(/ /,$line); # split it
          $markerinfo{$cols[1]}{CHR} = $cols[0];
          $markerinfo{$cols[1]}{POS} = $cols[$mapfilecol];

}
close(COLS);



#read ped file into memory
open(INPUT, "<$pedfile") || die("Failed to open $pedfile for reading");
my %sheep=();
my $scount=0;
while (<INPUT>)
{
          my $line = $_;
          chomp($line);
          $line =~ s/\n//g;  # \n
          $line =~ s/\r//g;  # \r
          $line =~ s/\t/ /g;

          #split col list
          my @cols = split(/ /,$line); # split it

          if ($cols[5] == $aff)  # lets just look in affecteds/controls
          {
                    $scount++;
                    for(my $i=6; $i<=$#cols; $i=$i+2)
                    {
                              $sheep{$cols[1]}{$i} = $cols[$i].$cols[$i+1];
                    }
          }
}
close(INPUT);
print "Finished reading ped - examining HZ in $scount individuals\n";


#now sort through data, one SNP at a time, looking for runs of 1 or more
#SNPs that are all HZ (or with wobble mismatch included)

#declare newvariables for this part
my @ids = sort {$a cmp $b} keys %sheep;
my @markers = sort {$a<=>$b} keys %{$sheep{$ids[0]}};
my $maxhz=0;
my $maxchrom=0;
my $hz=0;
my $oldc=-1;
my $gc=0;
my $hzstring="";
my $maxhzstring="";
foreach my $m (@markers) #check for all markers
{

          my $name = $datmap{$m};
          if (!exists($datmap{$m}))
          {
                    print "Ignoring m = $m\n";
          }
          else
          {

          #get details of this marker
          my $p = $markerinfo{$name}{"POS"};
          my $c = $markerinfo{$name}{"CHR"};
```

```
#check if on same chrom. If no, then check if we are in a run of HZ or not
if ($oldc ne $c)
{
          if ($hz>0)
          {
                    if ($hz>=$cutoff) #if hit end of chrom, and found a region of HZ, then save it

                              #must be the old chrom saved
                              $results{$hz}{$oldc}{$hzstring}=1;

                    }

                    $hz=0;
                    $hzstring="";
          }
}

$gc++;


#Get the genotypes for the sheep, seeing if HZ (get the first one listed in the ped file)
my $fail1=0;
my $a1 = "";
for (my $i=0; $i<=$#ids; $i++)
{
          if ($sheep{$ids[$i]}{$m} ne "00" && $sheep{$ids[$i]}{$m} ne "")
          {
                    $a1 = $sheep{$ids[$i]}{$m};
                    $i=$#ids+1;

          }
}

my $genostring="[";

#now complete with the rest, checking if the genotypes are HZ with the first
for (my $i=0; $i<=$#ids; $i++)
{
          $genostring = $genostring." $sheep{$ids[$i]}{$m}";


          if ($sheep{$ids[$i]}{$m} ne $a1 && $sheep{$ids[$i]}{$m} ne "00" && $sheep{$ids[$i]}{$m} ne "")
          {
                    $fail1++;
          }
}

$genostring = $genostring."]";

#if the number of 'failed' matches is less than our threshold, then save this HZ region
if ($fail1<=$mismatch &&  $a1 && $a1 ne "00" && $a1 ne "" && $a1 ne " ")
{

          #if saving, then lookup GFF details
          my @ks = sort {$start{$c}{$a} cmp $start{$c}{$b}} keys %{$start{$c}};

          my $found=0;
          my $genename="(no gene)";
          foreach my $k (@ks)
          {
                    #print "Checking if $cols[3] >= $k and $cols[3] <= $end{$chr}{$start{$chr}{$k}} \n";

                    if ($p >= $k && $p <= $end{$c}{$start{$c}{$k}})
                    {
                              $found++;
                              $genename="($names{$c}{$k}{$end{$c}{$start{$c}{$k}}})";
                    }
          }
```

```
                    $hz++;
                    $hzstring = $hzstring." $name($p)"."$genename"."$genostring";

          }
          else
          {
                    # this SNP failed our threshold test, but we need to check if this is the end of a run of HZ
                    if ($hz>0)
                    {
                              if ($hz>=$cutoff)
                              {

                                        #save old chrom
                                        $results{$hz}{$oldc}{$hzstring}=1;

                              }

                              $hz=0;
                              $hzstring="";
                    }

          }



          if ($m==$#markers)
          {
                    if ($hz>=$cutoff)
                    {
                              #save old chrom
                              $results{$hz}{$oldc}{$hzstring}=1;

                    }
          }

          #save old chrom
          $oldc = $c;

          }

}


# Now print all the saved regions, sorted from longest to shortest
print "Top results with <= $mismatch per genotype, reporting runs >= $cutoff\n\n";
print "HZcount\tCHROM\tMARKERS\n";

my @hzs = sort {$b<=>$a} keys %results;
foreach my $h (@hzs)
{
          my @chrs = sort {$b<=>$a} keys %{$results{$h}};
          foreach my $c (@chrs)
          {
                    my @annots = sort {$b<=>$a} keys %{$results{$h}{$c}};
                    foreach my $a (@annots)
                    {
                              print "$h\t$c\t$a\n";
                    }
          }

}
```

Appendix VI: perl source code for Haplotype_Extractor

Please note that font size has been decreased to assist with layout readability of the code

```
#!/usr/bin/perl

# Authors: Jez Supreme and Kim Carter, 2013
# Purpose: Script to extract a subset of markers from a set of pedigree files
#
# Inputs:  1. .dat text file, in standard linkage dat format
#          2. target markers (to extract), one marker per line
#          3. .ped text fle, in standard linkage ped format

# Outputs:  Subset of targeted SNPs, output in standard linkage ped and dat format

use strict;

#check correct arguments are supplied)
if ($#ARGV!=3) {
          print "Usage: Usage: Haplotpye_Extractor.pl <datfile.txt> <targetedSNPs.txt - 1 marker per line> <ped file.txt> <outputhaplotype.txt>\n";
          exit;
}


#declare variables
my $datfile = $ARGV[0];   ##full list of SNPs
my $exfile = $ARGV[1];   ##which SNPs targeted
my $pedfile = $ARGV[2];  ##pedigree file
my $newped = $ARGV[3];  ##output haplotype

my %map = ();

##### ASSUMPTION ####################

#ped file =
#
#   col 0 = FID
#       1 = IID
#       2 = FAID
#       3 = MOID
#       4 = SEX
#       5 = AFF
#       6 = SNP MARKER ALLELE 1
#       (and rest of SNP alleles)


#read DAT file into memory
open(DAT, "<$datfile") || die("Failed to open $datfile for reading");
my $pos=6;
while(<DAT>)
{
          my $line = $_;
          chomp($line);
          $line =~ s/\n//g;  # \n
          $line =~ s/\r//g;  # \r
        $line =~ s/\t/ /g;
```

```
          if ($line =~ m/^M/)  #markers only, ie skip A Affection
          {
                    my @cols = split(/ /,$line); # split it
                    $map{$pos} = $cols[1];
                    $pos = $pos+2;
          }
          elsif ($line =~ m/^A/)  #markers only, ie skip A Affection
        {
                my @cols = split(/ /,$line); # split it
                $map{5} = $cols[1];
        }
}
close(DAT);


#read extract markers file, one per line
my %toremove = ();
open(IN, "<$exfile") || die("Failed to open $exfile for reading");
while(<IN>)
{
          my $line = $_;
          chomp($line);
          $line =~ s/\n//g;  # \n
          $line =~ s/\r//g;  # \r
        $line =~ s/\t/ /g;

          $toremove{$line} = 1;
          print "Copying marker"
}
close(IN);


#read PED file, into memory, and extract only the subset of markers requested -> printing the new ped file
open(IN, "<$pedfile") || die("Failed to open $pedfile for reading");
open(OUT, ">$newped") || die("Failed to open $newped for writing");
open(OUTDAT, ">$newped.dat") || die("Failed to open $newped.dat for writing");
while(<IN>)
{
          my $line = $_;
          chomp($line);
          $line =~ s/\n//g;  # \n
          $line =~ s/\r//g;  # \r
        $line =~ s/\t/ /g;

          my @cols = split(/ /,$line); # split it
          print OUT "$cols[0] $cols[1] $cols[2] $cols[3] $cols[4] $cols[5]";
          for(my $i=6; $i<=$#cols; $i=$i+2)
          {
                    if (exists($toremove{$map{$i}}))
                    {
                              print OUT " ".$cols[$i]." ".$cols[$i+1];

                    }
                    else
                    {

                    }

          }
          print OUT "\n";
}
```

```
close(IN);


#print the new DAT file
my @keys = sort {$a<=>$b} keys %map;
foreach my $k (@keys)
{
          if ($k==5)
          {
                    print OUTDAT "A ".$map{$k}."\n";
          }
          else
          {
                    if (exists($toremove{$map{$k}}))
                {
                              #write to dat file
                              print OUTDAT "M ".$map{$k}."\n";
                    }
          }
}
close(OUTDAT);

#summary
print "\nNew haplotype written to $newped\n";
print "\nNew dat file written to $newped.dat\n";
```

Appendix VII: perl source code for the Plink_Parser

Please note that font size has been decreased to assist with layout readability of the code

```
#!/usr/bin/perl

# Authors: Jez Supreme and Kim Carter, 2013
# Purpose: Script to rear in Plink.model association outputs, and merge in nearest genes using genome locations from a GFF file
#
# Inputs:  1. Plink model file to parse
#          2. Formatted .map file in standard linkage map format (used to locate a marker)
#          3. Significant cutoff - maximum value to filter association p-values by, eg 0.05
#            4. Association model -  REC, GENO, DOM, TREND, ALLELIC

# Outputs:  Summary of plink model association, with nearest gene in a column separated format
#             SNPMARKER \t CHROM \t BP \t GENEID \t GENENAME \t GENEBP \t DISTANCETOGENE \t ALLELE1 \t ALLELE2 \t MODEL \t FRQ_AFF \t FRQ_UNAFF \t PVAL

use strict;

#GFF file location - Sheep genome version 3.1
my $gff = "/data/gff_files/Oarv3.1.protein.gene.gff3";
my $mapcols = 3;  #number of the basepair column in a linkage map file

#check required inputs are provided
if ($#ARGV!=3) {
          print "Usage: Usage: parse_plink_model.pl  <plink.model> <plink.map> <sig cutoff> <model>\n";
          exit;
}


#declare variables
my $sig = $ARGV[2];
my $model = $ARGV[3];
my %start = ();
my %end = ();
my %names = ();


#open GFF file, and read into memory
open(GFF, "<$gff") || die("Failed to open $gff for reading");
while(<GFF>)
{
          #trim newline character
          my $line = $_;
          chomp($line);

          $line =~ s/\n//g;  # \n
          $line =~ s/\r//g;  # \r

          my @cols = split(/\t/,$line); # split it

          if ($cols[2] eq "mRNA") # grab only MRNA (ie gene) features
          {
                    #scaffold004293     GLEAN     mRNA      16        1497      0.999942  -         .
          D=CCG000001.0;Name=Uncharacterized;Homology=B4AF47_BACPU;Note=BLAST.hit.in.refproteindatabaseB4AF47

                    #parse the name
```

```
                    my $namebit = $cols[8];
                    my @ns = split(/;/,$namebit); # split it
                    #print "@ns\n";

                    #parse the chromosome
                    my $chr = $cols[0];
                    #OAR11
                    if ($chr =~ m/OAR/)
                    {
                              $chr = substr($chr,3);
                    }


                    #store the start and end positions in hash tables
                    if (exists($start{$chr}{$cols[3]}))
                    {

                              #print "dupe\n";
                              $start{$chr}{$cols[3]} = $start{$chr}{$cols[3]}."::".$ns[1];
                    }
                    else
                    {
                              $start{$chr}{$cols[3]} = $ns[0];
                    }
                    $end{$chr}{$ns[0]} = $cols[4];

                    $names{$chr}{$cols[3]}{$cols[4]} = $ns[1];
          }
}
close(GFF);
#print "Finished reading GFF\n";


# Read the plink.model into memory
my %markers = ();
open(PLINK, "<$ARGV[0]") || die("Failed to open $ARGV[0] for reading");
while(<PLINK>)
{
          #CHR                          SNP   A1   A2     TEST            AFF          UNAFF        C
#HISQ   DF           P

          my $line = $_;
          chomp($line);

          #trim end of line characters and multiple whitespaces
          $line =~ s/\n//g;  # \n
          $line =~ s/\r//g;  # \r
        $line =~ s/\t/ /g;
          $line =~ s/\h+/ /g;

          $line =~ s/^\s+//;
          $line =~ s/\s+$//;

          if ($line eq "")
          {}
          else
          {
                    my @cols = split(/ /,$line); # split i
                    #print "$cols[1]::$cols[4]::$cols[9]\n";

                    #save this SNP result if its the correct model, and meets the significance cutoff
```

```
                    if ($cols[4] eq $model && $cols[9] ne "NA" && $cols[9] <= $sig)
                    {
                              $markers{$cols[1]} = $line;
                              #print "Storing marker $cols[1]: $line\n";

                    }
          }


}
close(PLINK);


#open the map file and check where our significnant markers are, relative to the GFF features
print "MARKER\tCHROM\tBP\tGENEID\tGENENAME\tGENEBP\tDISTANCETOGENE\tA1\tA2\tMODEL\tFRQ_AFF\tFRQ_UNAFF\tPVAL\n";
open(MAP, "<$ARGV[1]") || die("Failed to open $ARGV[1] for reading");
while(<MAP>)
{

          my $line = $_;
          chomp($line);

          #trim end of line
          $line =~ s/\n//g;  # \n
          $line =~ s/\r//g;  # \r
        $line =~ s/\t/ /g;
          $line =~ s/^\s+//;
          $line =~ s/\s+$//;


          my @cols = split(/ /,$line); # split it
          my $chr = $cols[0];
          if ($chr =~ m/^OAR/)
          {
                    $chr = substring($chr,3);
          }

          #found the marker in the significant list
          if (exists($markers{$cols[1]}))
          {
                    #print "here\n";

                    my @ms = split(/ /,$markers{$cols[1]});


                    my @ks = sort {$start{$chr}{$a} cmp $start{$chr}{$b}} keys %{$start{$chr}};  ##NEW

                    #check if the SNP falls within a gene
                    my $found=0;
                    foreach my $k (@ks)
                    {
                              #print "Checking if $cols[3] >= $k and $cols[3] <= $end{$chr}{$start{$chr}{$k}} \n";

                              if ($cols[$mapcols] >= $k && $cols[$mapcols] <= $end{$chr}{$start{$chr}{$k}})
                              {
                                        print
"$cols[1]\t$chr\t$cols[$mapcols]\t$start{$chr}{$k}\t$names{$chr}{$k}{$end{$chr}{$start{$chr}{$k}}}\t$k..$end{$chr}{$start{$chr}{$k}}\t\t$ms[2]\t$ms[3]\t$ms[4]\t$ms[5]\t$ms[6]\t$ms[9]\n";
                                        $found++;

                              }
                    }
```

```
                    #if a SNP doesn't fall within a gene, find the nearest upstream or downstream
                    if ($found==0)
                    {
                              #print "No gene found for $line\n";

                              my $close = 999999999;
                              my $close_gene = "\t\t\t\t\t\t\t\t\t\n";

                              my @starts = sort {$start{chr}{$a} cmp $start{$chr}{$b}} keys %{$start{$chr}};
                              foreach my $s (@starts)
                              {
                                        if ($s-$cols[$mapcols] <= $close && $s-$cols[$mapcols]>=0)
                                        {
                                                  $close = $s-$cols[$mapcols];
                                                  $close_gene =
"(Close_to)_".$start{$chr}{$s}."\t(Close_to)_".$names{$chr}{$s}{$end{$chr}{$start{$chr}{$s}}}."\t$s..$end{$chr}{$start{$chr}{$s}}\t$close\t$ms[2]\t$ms[3]\t$ms[4]\t$ms[5]\t$ms[6]\t$ms[9]\n";
                                        }

                                        if ($cols[$mapcols]-$end{$chr}{$start{$chr}{$s}} <= $close && $cols[$mapcols]-$end{$chr}{$start{$chr}{$s}}>=0)
                                        {
                                                  $close = $cols[$mapcols]-$end{$chr}{$start{$chr}{$s}};
                                                  $close_gene =
"(Close_to)_".$start{$chr}{$s}."\t(Close_to)_".$names{$chr}{$s}{$end{$chr}{$start{$chr}{$s}}}."\t$s..$end{$chr}{$start{$chr}{$s}}\t$close\t$ms[2]\t$ms[3]\t$ms[4]\t$ms[5]\t$ms[6]\t$ms[9]\n";
                                        }


                              }

                              print "$cols[1]\t$chr\t$cols[$mapcols]\t$close_gene";


                    }

          }
          else
          {
                    #die "Marker $cols[1] not found in GFF\n";
          }

}
close(MAP);
```